Exam 1, Spring 2022, SOLUTIONS

Problem #1

1. a. Let μ_1 - population mean salary for data scientists, μ_2 - population mean salary for software engineers.

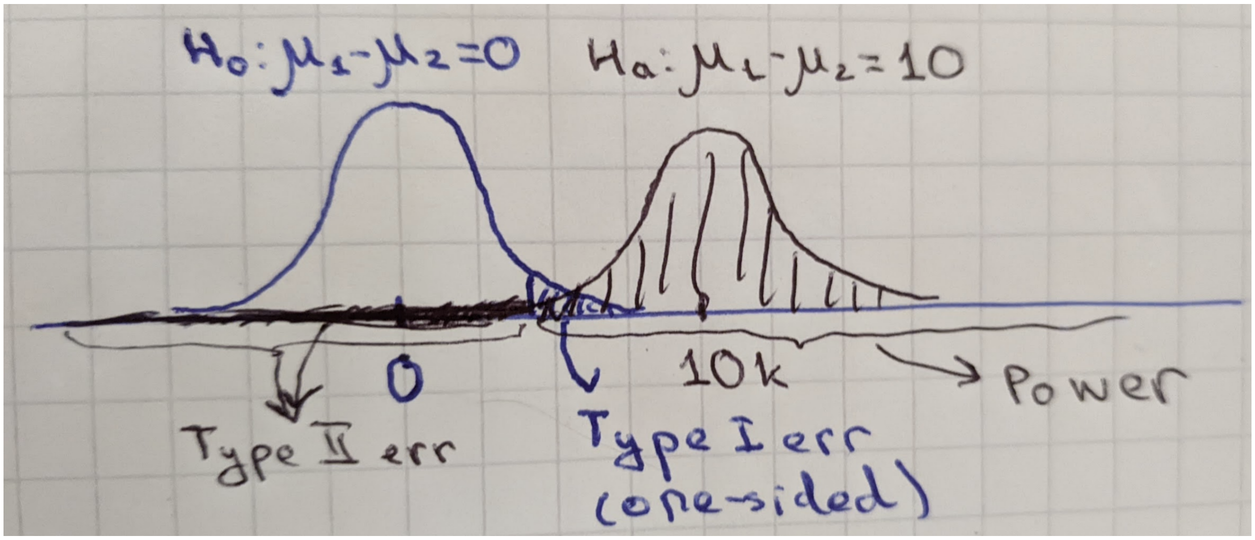
$$H_0\colon\ \mu_1 - \mu_2 = 0, \quad vs \quad H_a\colon\ \mu_1 - \mu_2 > 0$$

   b. Type I error = Reject H_0, given that H_0 is actually true. Type II error = Fail to Reject H_0, given that H_a is actually true (hence, H_0 is actually false).

   c. In this case, Type I error means concluding that the career change might be worthwile, while it actually won't be. Type II error means concluding that the career change won't be worthwile, while it actually will be.

2. a. $P(\text{Type I error}) = 0.05, Power = 1 - P(\text{Type II error}) = 0.75.$

   b. Picture below:



   c.
   - Decrease.
   - Decrease.
   - Decrease.
   - Decrease.

3. a. Extremely analogous to the HW problem about comparing 2 proportions, bar the "worst-case scenario" part. Just need to solve for n, remembering that $n = n_1 = n_2$ due to equal sample sizes.

   b. Come up with a reasonable range of values for those salaries $(min - max)$, assume that it's approximately normally distributed, divide by 6, recalling that by empirical rule $\approx$ 100% of a normally distributed quantity is contained within 3 standard deviations of its mean.

Problem #2

1. χ^2 test. Hypotheses can be formulated in a couple of ways:

   H_0: {Job satisfaction is independent of job type}, vs H_a: {Job satisfaction is dependent on the job type}

   or

   H_0: {Job satisfaction is the same for DS and SE}, vs H_a: {Job satisfaction differs between DS and SE}

2. O stands for **o**bserved cell count. E stands for **e**xpected cell count **under H_0 hypothesis of independence**. "i" is the row index, "j" is the column index, hence we are summing over all cells (all level combinations across two factor variables).

3. Cell $(1,2)$: $\frac{200\times 210}{600} = 70$. Cell $(2,3)$: $\frac{180\times 400}{600} = 120$. Both are considerably above the observed values of 50 and 90, respectively, which serves as evidence against H_0, because what we observe doesn't match what would be expected under H_0

4. Theoretical distribution should be $\chi^2_{(2-1)\times(4-1)}$, ore χ^2_3. For the shape, see https://istats.shinyapps.io/ChisqDist/ and recall the in-class explanation from where the the step-by-step χ^2-test procedure was provided (marking the tail to the right of the observed X^2-statistic value).

5. Due to tiny p-value, we reject the null hypothesis of independence of between job satisfaction and job type. Using "Happy" category, we see $P(\text{Happy} \mid \text{Dat Sci}) = 90/200 = 0.45$, $P(\text{Happy} \mid \text{Soft Eng}) = 90/400 = 0.225$, so one is twice as likely to be happy when working as a data scientist as opposed to software engineering.

6. P-value would have likely increased (see the slide with 3 tables, talking about increasing sample size from 100 to 200 to $10,000$), pointing to less statistical significance. Practical difference, on the other hand, would remain the same (e.g. $P(\text{Happy} \mid \text{Dat Sci}) = 9/20 = 0.45$, $P(\text{Happy} \mid \text{Soft Eng}) = 9/40 = 0.225$, still the same). That shows the importance of distinguishing between statistical and practical significance in χ^2 test, indicating that X^2-statistic value and p-value don't quantify the **strength** of the relationship, but solely point to the existence of the relationship.

Problem #3

1. $medv_i = \beta_0 + \beta_1 age_i + \epsilon_i,\ \epsilon_i \sim_{i.i.d.} N(0,\sigma^2)$,

   β_1 describes the population parameter value (slope of the $medv \sim age$ line if could fit it to **the entire population**), while $\hat{\beta}_1$ is the sample estimate (slope of the $medv \sim age$ line fitted just to **that sample**).

2. If we were to take many random samples from the population, and calculate $\hat{\beta}_1$ estimate for each of those samples, on average it will be equal to population value β_1.

3. By how much, on average, does the sample estimate $\hat{\beta}_1$ deviate from population parameter value β_1.

   OR

   The standard deviation of sampling distribution of $\hat{\beta}_1$ (if we were to take many random samples and calculate $\hat{\beta}_1$ estimate for each sample, how spread out will the $\hat{\beta}_1$ values be).

4. $\widehat{medv} = 30.97 - 0.12 \times age$.

5. $RSE = 8.527$ - our predicted median prices miss the true prices by $8.527 \times 1,000\$ = 8,527\$$, on average.

   $R^2 = 0.14$ - our model (linear regression with age as predictor) explains 14% of variability in the median house prices.

6. We are 95% confident that, per 1%-point increase in % of old buildings in the neighborhood, median house price will drop off by between 97$ to 150$, on average.

7. - For neighborhoods with 50% of older buildings, the median house price will be $24,820$$, on average.
   - We are 95% confident that, for neighborhoods that have 50% of older buildings, the **average** median house price will be between $23,930$$ and $25,710$$.
   - We are 95% confident that, for neighborhoods that have 50% of older buildings, any median house price is going to be between $23,930$$ and $25,710$$. (Alternative interpetation: "95% of **all** median house prices for neighborhoods that have 50% of older buildings will be between $8,040$$ and $41,600$$").