Exam 3 review (draft: 2019/05/02-20:04:41)

Exam 3 covers all material since the 2nd exam. You will still need to remember much of the material since the first day of class though, including, but not limited to, many of the rules of probability.

My recommendation is to study the two quizzes, make sure you understand everything on those perfectly. Make sure you have looked over any example problems included in the course notes that I have posted online (for chapters 5,6,7). Make sure you understand the webwork homework problems. (also note that the chapter numbers for the webwork assignments do not necessarily align with the chapter numbers for my notes). Of course, you should also understand everything and all examples covered in class. Sometimes these vary somewhat form what is seen on homework or in my online notes. You should also pay attention to what we have done with R.

Here is a summary list of topics (note that this list is not completely exhaustive):

- Confidence intervals
- Hypothesis testing
- Linear regression

Example/practice problems:

1. For a hypothesis test of a single mean calculate the p-value in each case below for all three possible tests: 2-sided, and both 1-sided tests. Write the R code.

   (a) Test statistic $= 1.5$, $n = 10$, σ known, normal population.
   Solution: A test statistic is the sample mean minus the null hypothesis mean, so our sample mean is above the null hypothesis mean. Thus, we generally have no reason to use the 1-sided alternative $H_a : \mu < \mu_0$ as it will give a p-value above 50%.
   Population normal, known variance, so we use a z test statistic, so p-values are:
   $H_a : \mu > \mu_0$, `1-pnorm(1.5)` or `pnorm(-1.5)`
   $H_a : \mu \neq \mu_0$, `2*pnorm(-1.5)` or `2*(1-pnorm(1.5))`

   (b) Test statistic $= 1.5$, $n = 10$, σ unknown, normal population.
   Solution: n small, unknown variance, so we use a t test statistic, so p-values are:
   $H_a : \mu > \mu_0$, `1-pt(1.5,9)` or `pt(-1.5,9)`
   $H_a : \mu \neq \mu_0$, `2*pt(-1.5,9)` or `2*(1-pt(1.5,9))`
   Again, we don't really need to test with alternative $H_a : \mu < \mu_0$ since we automatically know the p-value will be above 50%.

   (c) Test statistic $= 1.5$, $n = 50$, σ unknown, normal population.
   Solution: n large, by CLT, we use a z test statistic, so p-values are:
   $H_a : \mu > \mu_0$, `1-pnorm(1.5)` or `pnorm(-1.5)`
   $H_a : \mu \neq \mu_0$, `2*pnorm(-1.5)` or `2*(1-pnorm(1.5))`

   (d) Test statistic $= 1.5$, $n = 50$, population distribution unknown.
   Solution:
   n large, by CLT, we use a z test statistic, so p-values are as in (a) and (c).

   Note that in all cases, a test statistic of 1.5 will give a fairly large p-value. The smallest possible p value would be a z test statistic. We know that $z = \pm 2$ gives 2.5% tails, so our smallest possible p value is larger than 2.5%. The smallest p-value is precisely `pnorm(-1.5)`≈ 0.0668072, so we would actually not reject H_0 in any of these cases.

2. If a sample proportion of successes is 0.80 out of 50 trials, estimate the population proportion with a 95% confidence interval.

   Solution:

   For proportions with greater than 5 successes and failures, we use the CLT approximation.

   $\hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

   I'll use the 68-95-99.7 rule that ± 2 standard deviation si approximately 95% for the standard normal: $0.8 \pm 2\sqrt{\frac{0.80(0.2)}{50}} \approx (0.6868629, 0.9131371)$

3. For the previous problem, test the hypothesis that the true proportion is below 70% at $\alpha = 0.05$.

   Solution:

   Out z test statistic is

   $z = \frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)n}} = \frac{0.8-0.7}{\sqrt{0.7(1-0.3)50}} \approx 2.886751$

   We should already be able to tell that this is well below the 5% level. The p-value is `pnorm(-z)`≈ 0.001946209. We reject the null hypothesis. Our sample proportion is a very rare high value under the assumption that the null hypothesis is true.

4. If we test a hypothesis at significance level α, and we find that $p > \alpha$, what do we conclude?

   Solution:

   We do not reject the null hypothesis. The p value represents how rare our data is under the assumption that the null hypothesis is true. A p value above our chosen level of significance means that our data is not sufficiently rare in our opinion.

5. You are given that for paired data (x, y), the sample mean of X is 6, the sample variance of X is 1.2, the sample mean of Y is 38, and the sample variance of Y is 20, and their covariance is 3. Find the least-squares linear regression line for $y = a + bx$. Calculate everything by hand without any R code.

   Solution:

   The slope is $b = \frac{cov(x,y)}{var(x)} = \frac{3}{1.2} = 2.5$ and the intercept is $a = \overline{y} - b\overline{x} = 38 - 2.5 \cdot 6 = 38 - 15 = 23$ So $Y = 23 + 2.5X$.

6. Draw an example scatterplot of a paired dataset that has correlation coefficient:

   - near zero
   - near 1
   - near -1
   - not close to 1 but still positive
   - not close to -1 but still negative

   Solution:

   I don't expect you to be able to reproduce this R code, but this will allow to to make many scatterplots and check the correlation and how linear they look.

```
> sig=1
  x=rnorm(100)
  y=x+rnorm(100,0,sig)
  plot(x,y)
  cor(x,y)sig=1
  x=rnorm(100)
  y=x+rnorm(100,0,sig)
  plot(x,y)
  cor(x,y)
```

Solution:

Run this code for different values of `sig` $= \sigma_\epsilon$.

$\sigma_\epsilon = 10$ given correlation coefficient ρ that jumps around a lot, usually between -0.25 and 0.25, so such data would be quite uncorrelated.

$\sigma_\epsilon = 1$ given correlation coefficient $\rho \approx 0.7$ mostly, usually between 0.6 and 0.8.

$\sigma_\epsilon = 0.5$ given correlation coefficient ρ usually around 0.9.

$\sigma_\epsilon = 0.1$ given correlation coefficient ρ very close to 1.

To get negative correlations, change the 3rd line to `y=-x+rnorm(100,0,sig)`.