ps3 — variability

questions appended with "i" are optional

I — asymptotics: large sample sizes

By now you should be familiar with the concept of **variance** as a measure of spread of a distribution:

$$\mathrm{Var}[X] = \mathrm{E}\big[(X - \mathrm{E}[X])^2\big]$$

or, writtten as a sum:

$$\sigma^2 = \frac{1}{n}\sum_i^n (X_i - \bar{X})^2 \qquad\qquad s^2 = \frac{1}{n-1}\sum_i^n (X_i - \bar{X})^2$$

i.e. variance is the expected value (i.e. the average) of the squared deviations from the mean. Note σ^2 denotes the population variance and s^2 denotes the sample variance.

1. Show that $\mathrm{E}\big[(X - \mathrm{E}[X])^2\big]$ can also be expressed $\mathrm{E}[X^2] - \mathrm{E}[X]^2$. *Hint:* expand the brackets and recognize that $\mathrm{E}[X]$ is a constant.

Consider now the following scenario: let X be a random variable that follows some distribution. Suppose we observe n instances of X, i.e. we have a collection of n RVs, $X_1, ..., X_n$, each of which is independent and has same distribution as X (i.e. the RVs are i.i.d.). Suppose we define another random variable S_n, where $S_n = \sum_i^n X_i$, such that $\bar{X}_n = \frac{1}{n} S_n$.

2. What do we need to assume about $\mathrm{E}[X]$ to conclude that $\bar{X}_n \to \mu$ as $n \to \infty$?

3. Suppose X can only take two values, -1 and 1. What are the lowest and highest possible values of S_n?

4. Using the same definition of X in q3, suppose we've made ten measurements of X, i.e. $n = 10$, and we have a sample mean $\bar{X}_{10} = 0.2$. We then add a new measurement X_{11} to the sample, where X_{11} is i.i.d. with X. What are the lowest and highest possible values for $\bar{X}_{11}$?

5. Now suppose we've made *a hundred* measurements of X, i.e. $n = 100$, and we have a sample mean $\bar{X}_{100} = 0.02$. We then add a new measurement X_{101} to the sample, where X_{101} is i.i.d. with X. What are the lowest and highest possible values for $\bar{X}_{101}$? Compare how the variability in $\bar{X}$ changes between q4 and q5.

The question(s) above should have demonstrated to you an example of *convergence*. This is an important idea in statistics—below we'll introduce two convergence theorems that will come up a lot in this course (and everywhere).

The **law of large numbers** (specifically, the *strong* law of large numbers) says the sample mean converges to its expected value when n is large enough:

$$\bar{X} \longrightarrow \mathrm{E}[X] \qquad n \longrightarrow \infty$$

6. Zenith and Quasar are beset by the problem of having to decide who does the dishes on a particular day. Quasar proposes they roll a six-sided die, and that if the die returns a value higher than 3, Quasar does the dishes, and if it returns a value smaller than 3, Zenith does the dishes. Let X be the RV for the outcome of the die on a particular day, i.e. $X \in \{1, ..., 6\}$. Let Q be the RV for whether Quasar does the dishes on a particular day, i.e. $Q \in \{0, 1\}$. What is $\mathrm{E}[X]$? What is $\mathrm{E}[Q]$? The proportion of times Quasar does the dishes in n days is $\bar{Q}_n$. What is the long run value of $\mathrm{E}[\bar{Q}_n]$? Is this system fair?

7. Zenith then proposes a new system: that they should keep track of the *average* of the die scores, $\bar{X}_n$, where $\bar{X}_n = \sum_i^n X_i$, and that Quasar should only do the dishes if the *average* is greater than 3. Now what is the long run value of $\mathrm{E}[\bar{Q}_n]$? What does this mean for Quasar?

8. Suppose you find yourself in a decadent game at a casino, where you bet some money X then spin a wheel. If you win, you get X dollars, and if not, you lose X dollars. You decide the best strategy is to bet X_1 on the first round, and if you lose, to go double or nothing and bet $X_2 = 2X_1$ on the second round. Suppose you are somewhat the worse for drink, and keep repeating this strategy, doubling your bet each round, until you win (or go bankrupt). If you win on the first round, your payoff is X_1. If you win on the nth round, your payoff is $2^n X_1$. Note this sequence of payoffs/losses doesn't satisfy the assumptions of the LLN. Which assumption(s) does it violate?

9. Historically, a basketball player has tended to make one basket for every two shots taken. During a particular game there is a period when she makes five baskets in a row. The team then starts to send her the ball more often, to reap the rewards of this "lucky streak". Is this a good strategy? Explain your reasoning.

The **central limit theorem** says that if $X_1, ..., X_n$ are i.i.d random variables, and n is large enough, the distribution of the sample mean becomes approximately normal, with mean $\mathrm{E}[X]$ and variance $\frac{\mathrm{Var}[X]}{n}$:

$$\bar{X} \sim \mathcal{N}\left(\mathrm{E}[X], \frac{\mathrm{Var}[X]}{n}\right)$$

10. This question is to help you understand the idea of a **sampling distribution**. Let X be a RV that follows a continuous uniform distribution between 0 and 1, i.e. $X \sim \mathcal{U}(0,1)$. Sketch the pdf of X. What is $\mathrm{E}[X]$ and $\mathrm{Var}[X]$?

    Suppose now you observe n i.i.d instances of X, where $X_1, ..., X_n \sim \mathcal{U}(0,1)$. Let $\bar{X}_n = \sum_i^n X_i$. Since we are treating $\bar{X}_n$ as a random variable, it has a distribution—it's known as the *sampling distribution of the mean*. Using the CLT, can you think of a statement describing the distribution of $\bar{X}_n$? Simplify as much as you can.

    In R, simulate the distribution of $\bar{X}_n$ by drawing many random samples of X and plotting the sampling distribution of $\bar{X}_n$. Use different sample sizes, $n = 1, 5, 15, 30, 100....$ For each sampling distribution, compute $\mathrm{E}[\bar{X}_n]$ and $\mathrm{Var}[\bar{X}_n]$, and check that the values agree with your theoretical calculations. What do you notice about the sampling distribution of $\bar{X}_n$ as n increases? Note the `runif()` function is useful for drawing random numbers from a uniform distribution. For tips on generating random numbers in R, check out this link or the notes in chapter 10.

For q10, you don't need to present your code or any of its output—just state and interpret your results, as applicable. If you dare, repeat the simulation with a different underlying distribution (e.g. try the exponential or Poisson). This exercise is for you to convince yourself that the sampling distribution of the mean will converge to a normal distribution for large enough n—this is a very important result, as it implies many random processes can be modeled by a normal distribution, even ones that follow different distributions.

11. Recall that if $Z \sim \mathcal{N}(\mu, \sigma^2)$ then $\mathrm{P}(|Z - \mu| < 2\sigma) = 0.95$. How large does n have to be so that $\bar{X}$ has 2.5% probability of being greater than $\mu + \frac{1}{10}$? Simplify.

12. Suppose that the height of giraffes in Cyprus has a mean of 12 meters and a standard deviation of 1.2 meters. You draw 40 giraffes at random. Stating any assumptions you make, use the CLT to find the approximate probability that the average height of giraffes in your sample is at least 11.8 metres.

13. Suppose we have a factory that produces lightbulbs and that each bulb independently has a 5% probability of dying after 4 years. Suppose that during a particular year we produce 1250 bulbs. Let X be the RV for the number of bulbs still working after 4 years. Use the CLT to approximate $\mathrm{P}(X > 1200)$.

14*i*. Suppose we have a factory that produces brake pads, and that there are on average 2 faulty brake pads per 400 produced. Suppose we sell 4000 brake pads. Use the CLT to approximate the probability that fewer than 20 of the brake pads sold are faulty.

15*i*. On day 1 a stock price has a value of zero. Let U be the RV for how much the stock price changes relative to the previous day. Suppose U can only take two values: $U = 1$ if the price increases by one dollar, and $U = -1$ if it decreases by one dollar. Suppose that $\mathrm{P}(U = 1) = \mathrm{P}(U = -1) = \frac{1}{2}$. Now, let X_n be the value of the stock on day n, i.e. $X_n = \sum_i^n U_i$. What is $\mathrm{E}[X_n]$ and $\mathrm{Var}[X_n]$? Note this is known as a **random walk**. Simulate X_n and plot X_n versus n for different values of n. If you have time, repeat the simulation several times—you'll notice that each run looks different, even though they were generated by the same underlying process. How do your theorized values for $\mathrm{E}[X_n]$ and $\mathrm{Var}[X_n]$ explain your observations?

II — intervals and tests

Suppose that a notorious European car manufacturer produces a car called the Flog. You, an independent reviewer, decide to test how its emissions of nitrogen dioxide vary in different environments.

Suppose you make 75 measurements of the Flog's NO_2 emissions on a rolling road (a controlled testing apparatus), and 75 measurements of its NO_2 emissions on the open road. The dataset `emissions.csv` contains your observations for each group, in units of mg/m^3.

The following questions will get you to use this data to compute some confidence intervals and perform some basic hypothesis tests. The file `ps3.Rmd` has most of the code required to answer these questions—you can either use it or write your own. Note you don't need to turn in any code, just your interpretations of the results.

1. What is the sample mean and sample standard deviation of NO_2 emissions in each group?

2. Compute a 95% confidence interval for each group.

3. Suppose you define a new random variable D, for the difference in emissions between the two groups. Assuming the RVs are independent, what is $\mathrm{E}[D]$ and $\mathrm{Var}[D]$?

4. Compute a 95% confidence interval for the mean difference in emissions. Does this interval contain zero?

5. Suppose you repeated the experiment after collecting a new sample of data, and you computed a new 95% confidence interval for the mean difference. Which of the following might be different in this new experiment: (1) population means, (2) population variances, (3) sample means, (4) sample variances, (5) the center of the CI, (6) the length of the CI, (7) whether or not the CI contains zero, (8) whether or not the CI contains the difference in population means.

6. Suppose you want to perform a hypothesis test to determine whether there is a really a difference in emissions between the two groups. State the null and alternative hypotheses for testing equality in means between the two groups. Perform a two sample t-test. State and interpret the p-value. What do you conclude at the 5% significance level?

7. Now compute a 99% confidence interval for the mean difference. Does this interval contain zero? If you had performed the test at the 1% level, what would you have concluded?

Next, consider the following scenario: a randomized experiment is conducted to assess the effectiveness of four drugs at reducing nausea after an operation. Below are the findings:

	incidence of nausea	number of patients
placebo	46	88
drug 1	24	75
drug 2	15	81
drug 3	41	90
drug 4	23	66

In this study we're interested in finding whether each of the four drugs has a demonstrable effect on the incidence of nausea following an operation. In each case we compare the drug data to the placebo data. Naturally, if the drug has no effect, then the drug data should *resemble* the control group (the placebo), and vice versa.

8. Based on the data in the table, what is the expected proportion of nausea incidence when no drugs are taken?

9. The **Chi-squared test** is useful for testing whether groups of categorical data resemble each other. In this example, we can use the Chi-squared test to determine whether each of the drugs has an effect on nausea incidence, by making a comparison to the placebo group. The null hypothesis for this test is that nausea incidence is *independent* of the drug (i.e. the drug has no effect).

   Test each of the four drugs versus the placebo at the 5% level. Use the Chi-squared test—you can find the code for this in `ps3.Rmd`. Comment on your findings.

In a hypothesis test, the null is assumed until there is strong evidence to suggest we should reject it. Of course, the outcome of a single hypothesis test does not necessarily lead to the correct result—there are two kinds of error we can make:

- **Type 1 Error:** rejecting the null, when the null is *actually* true. The type 1 error should be no more than the significance level of a test.
- **Type 2 Error:** failing to reject the null, when the null is *actually* false.

Usually the foremost goal in hypothesis testing is to minimize the likelihood of type 1 error. Another useful quantity is:

- **Power:** the probability of *correctly* rejecting a false null. Note this is the complement of Type 2 Error, i.e. Power = 1 - T2E.

The secondary goal in hypothesis testing is to choose the most *powerful* test (i.e. the test with the smallest likelihood of type 2 error).

10. Suppose we conduct a test and the probability we make a type 1 error is 5%. We then do the same test again using a new independent sample of data. What is the probability that both tests correctly fail to reject the null?

11. Suppose we conduct the same test n times, with a new sample of data each time. If each test has a 5% probability of making a type 1 error, what is the probability that we make no type 1 errors in any of the n independent tests?

12. What kind of random variable could we use to model the number of type 1 errors out of n independent tests?

13. Suppose each of our n tests has 60% power, meaning that the probability of a type 2 error is 40%. In the long run, the proportion of tests that correctly reject the null hypothesis converges to some value. Why? And: what is this value ?

In the file `ps3.Rmd`, under the header "exponential distribution", we have written a function that generates a sample of n observations from the exponential distribution with true mean $\mu = 2$. The function then compares the sample against the null hypothesis that $\mu = 1$, using a two-sample t-test for a difference. If the null is rejected it returns `TRUE`, and if not it returns `FALSE`.

14. Run the experiment once with $n = 30$ and true mean $\mu = 2$. What do you get? What does it mean?

15. Now, repeat the experiment 1000 times and report the proportion of times the null is rejected. What does this value represent?

16. Do the same as in q15, but change the sample size to $n = 40$. Is the proportion of rejections different? Explain your results.

17. Change the parameters such that $n = 30$ and the true mean is $\mu = 1.5$. What do you notice? Explain your results.

18. Without running any more code, what do you expect the proportion to be if we changed the true mean μ to be 1? Why?

19. Set the parameters to $n = 30$ and $\mu = 1.1$. Using trial and error, find the value of n large enough so that the power of the test is 80%. Comment on your findings.

III — joint variability

Lastly, we will introduce some concepts from joint variability—the tendency for variables to vary *together* (i.e. exhibit dependency) .

Covariance is a measure of the extent to which two variables vary together in the same direction. Formally, the covariance of two variables X and Y is defined:

$$\text{Cov}[X, Y] = \text{E}\big[(X - \text{E}[X])(Y - \text{E}[Y])\big]$$

or, written as a sum:

$$s_{XY} = \frac{1}{n-1}\sum_{i}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

1. In the last problem set we told you that variance is additive for *independent* RVs, i.e. $\text{Var}[X+Y] = \text{Var}[X] + \text{Var}[Y]$ provided X and Y are independent. We'll now consider the case that X and Y are are dependent.

   Using the fact that $\text{Var}[X] = \text{E}[X^2] - \text{E}[X]^2$, show that the full expression for $\text{Var}[X+Y]$ is:

   $$\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y] + 2\text{Cov}[X, Y]$$

   Note that if two variables are independent, then $\text{E}[XY] = \text{E}[X]\text{E}[Y]$ (the multiplication rule). Use this to show that the last term in the expression for $\text{Var}[X+Y]$ vanishes if X and Y are independent.

2. Suppose $U \sim \text{Bin}(n, p)$ and $V \sim \text{Bin}(m, p)$ are independent binomial RVs. What are $\text{E}[U+V]$ and $\text{Var}[U+V]$?

3*i*. Consider a random variable X, whose value is determined as follows: if we flip a coin and it lands heads, we let $X \sim \mathcal{U}(0, 1)$, and if it lands tails, we let $X \sim \mathcal{U}(3, 4)$. Find $\text{E}[X]$ and $\text{Var}[X]$.