Solutions to Sample Final Examination

Math 125 *Kovitz* Spring 2008

1. (a) Option (iii) is right: regression effect (regression is toward the mean).

   (b) The 96th percentile has 96% of the area below it. That makes the upper tail 4%, to find a balanced area in the middle, remove the equal lower tail of 4% from the percentile rank to leave 92%.

   That 92% gives a z of 1.75.

   In standard units, his math SAT score was 1.75. The regression prediction for his verbal SAT score is $0.4 \times 1.75 = 0.7$ in standard units. This has about 52% of the area in the middle and $\frac{100\%-52\%}{2} = 24\%$ in each tail. The percentile rank will consist of the middle area and the lower tail which is $52\% + 24\% = 76\%$.

   (c) The 76th percentile verbal SAT score was about 0.7 in standard units, as we just found. The regression prediction for his math SAT score is $0.4 \times 0.7 = 0.28 \approx 0.3$ in standard units. This has about 24% of the area in the middle and $\frac{100\%-24\%}{2} = 38\%$ in each tail. The percentile rank will consist of the middle area and the lower tail which is $24\% + 38\% = 62\%$.

      i. False. That's the regression fallacy. In each case there is regression toward the mean. The second time, we use the other regression line and move closer to the mean, rather than getting back where we started.

      The only time that the two regressions get you back to the original place is when there is perfect correlation. In all other cases additional uncertainty is introduced in the second regression, bringing the ultimate answer closer to the mean than where we started originally.

2. (a) The averages are 3 and 4, respectively; the SDs are 1 and 2, respectively. The average product of the values in standard units is then

$$\frac{(0 \times 0) + (0 \times -1) + (-2 \times -1.5) + (0 \times 0.5) + (1 \times 0.5) + (1 \times 1.5)}{6}$$
$$= \frac{5}{6} \approx 0.83.$$

$r = 0.83$.

The slope of the regression line is

$$r \times \frac{\text{SD of } y}{\text{SD of } x} = \frac{5}{6} \times \frac{2}{1} = \frac{5}{3}.$$

The equation of the regression line is

$$y = mx + b = \frac{5}{3}x + b.$$

To find b, substitute the point of averages, $(3, 4)$, into the equation, getting $4 = \frac{5}{3} \times 3 + b$ and $4 = 5 + b$; so $b = -1$.

Equation: $y = \frac{5}{3}x - 1$.

When $x = 1.5$, $y = \frac{5}{3} \times 1.5 - 1 = 2.5 - 1 = 1.5$.

So, when $x = 1.5$, $y = 1.5$.

(b) i. The equation of the regression line is of the form $y = mx + b$ and given the slope is 6, the equation will be $y = 6x + b$. Substitute the point of averages into the equation to get $152 = 6(67) + b$. Then $152 = 402 + b$ and $b = -250$.

The equation is $y = 6x - 250$. Plugging in 72 for x gives $y = 6(72) - 250 = 432 - 250 = 182$ pounds.

An easier way is to note that the 72 inch man is 5 inches above the average of 67 inches. The slope of the line is 6 pounds per inch. That means for the regression line, which predicts 152 pounds for a 67-inch-tall man, every additional inch in height will increase the prediction of the weight by 6 pounds. This person, being 5 inches inches above 67 in height, will be 6×5 pounds, or 30 pounds heavier than 152 pounds. That means that the prediction is 182 pounds.

Or one could recapture r by noting that $r \times \frac{\text{SD of } y}{\text{SD of } x} = 6$. Then $r \times \frac{30}{4} = r \times 7.5 = 6$. Solving this gives $r = 6/7.5 = 0.8$.

Then predicted $y = 152 + \left(\frac{72-67}{4} \times 0.8 \times 30\right) = 152 + 1.25 \times 0.8 \times 30 = 182$ pounds.

ii. The standard units for 72.4 is $\frac{72.4-67}{4} = 1.35$.

The area under the normal curve from -1.35 to $+1.35$ is about 82%. Each tail is about $\frac{100-82}{2} = 9\%$. Add the lower tail to the middle area to get 91%, the percentile rank.

iii. If the percentile rank is 40%, the left tail is 40%. The corresponding right tail will also be 40%. That leaves $100\% - 40\% - 40\% = 20\%$ in the middle. Look up that area to find the standard units of $z = 0.25$.

In this case, the 40th percentile is below average. Use -0.25.

So $-0.25 = \frac{x-67}{4}$, $-1 = x - 67$, and $x = 67 - 1 = 66$ inches.

3. (a) with, independent.
   (b) without, mutually exclusive.
   (c) without, mutually exclusive.
   (d) with, mutually exclusive.
   (e) with, independent.
   (f) without, independent.

4. (a) The exact chance is $\frac{40!}{20!\times 20!}(.5)^{20}(.5)^{20} \approx 12.54\%$.
   Use the binomial formula to find the chance that an event with chance 1/2 will occur exactly 20 times out of 40. Here $n = 40$, $k = 20$, $n - k = 20$, $p = 0.5$, and $1 - p = 1 - 0.5 = 0.5$.

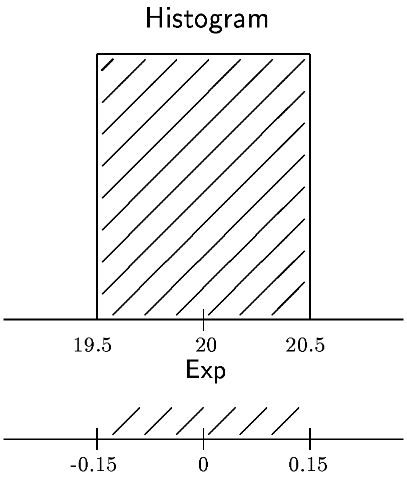
   (b) The expected number of heads is 20; the SE is 3.1623. Since the SD of a zero-one box with one 0 and one 1 is 0.50, and the SE is the product of the square root of the number of draws and the SE, we get $\sqrt{40} \times 0.50 = 6.3246 \times 0.50 = 3.1623$.

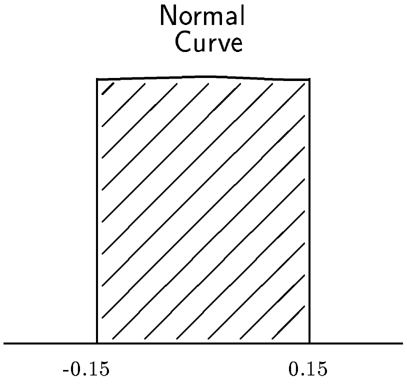
   We seek the area of the rectangle over 20, and need to find the standard units for its endpoints, 19.5 and 20.5.

   That gives us $\dfrac{19.5 - 20}{3.1623} = -0.158$ and $\dfrac{20.5 - 20}{3.1623} = 0.158$.

   For our table, round them to the nearest 0.05, getting +0.15 and −0.15.

   (The formula: std. units for the sum of draws $= \dfrac{\text{endpoint} - \text{EV of the sum}}{\text{SE for the sum}}$.)





Chance ≈ shaded area
≈ 11.92% (See comment below.)

Comment: the exact chance is ≈ 12.54%.

With interpolation, a better estimate for the area of the block is $11.92\% + \frac{0.008\%}{.05\%} \times (15.85\% - 11.92\%) = 11.92\% + 0.6288\% = 12.5488\%$, a really close approximation.

When doing the approximation, the curve doesn't curve much. The area under the normal curve is nearly rectangular, so there's not much error.

5. This is like 420 draws from the box $\boxed{\boxed{1}\ \boxed{2}\ \boxed{3}\ \boxed{4}\ \boxed{5}\ \boxed{6}}$. The average of this box is 3.5 and the SD of this box is 1.7 (calculating the SD with the method of Chapter 4 and getting $\sqrt{\frac{(1-3.5)^2+(2-3.5)^2+(3-3.5)^2+(4-3.5)^2+(5-3.5)^2+(6-3.5)^2}{6}} = \sqrt{\frac{17.5}{6}} \approx 1.7$).

   The expected value of the total spots on 420 rolls is $420 \times 3.5 = 1470$.

   The chance error in 1407, the observed total, is found by subtracting the expected value from the observation, getting $1407 - 1470 = -63$.

   The SE for the total of the spots on 420 rolls is $\sqrt{420} \times 1.7 = 35$.
   The standard units for the observed sum is calculated by

$$\text{std units for sum} = \frac{\text{observed sum} - \text{expected value of the sum}}{\text{standard error of the sum}} = \frac{1407 - 1470}{35} = \frac{-63}{35} = -1.8 \text{ standard units.}$$

   It's negative because the observed sum is less than the expected value of the sum.

6. The first thing to do it to set up a box model. There should be 30,000 tickets in the box, one for each registered voter; 12,000 are marked 1 (Democrat) and 18,000 are marked 0. The number of Democrats in the sample is like the sum of 1,000 draws from the box. The fraction of 1's in the box is 0.4. The expected value for the sum is $1{,}000 \times 0.4 = 400$. The SD of the box is $\sqrt{0.4 \times 0.6} \approx 0.49$. The SE for the sum is $\sqrt{1000} \times 0.49 \approx 15.5$.

   (a) The expected value for the percent is 400 out of 1,000, or 40%. (Or note: EV equals percentage of Democrats in the town: $12{,}000/30.000 \times 100\% = 40\%$.) The SE for the percent is 15.5 out of 1,000 or 1.55%.
   The SE for the percent may also be calculated directly from the SD of the box and the sample size by

$$\frac{\text{SD of the box}}{\sqrt{n}} \times 100\% = \frac{0.49}{\sqrt{1{,}000}} \times 100\% = 1.55\%.$$

   (No surprise about the expected value: 40% of the registered voters are Democrats.)

   (b) The percentage of Democrats in the sample will be about 40.0%, give or take 1.55% or so. Parts (a) and (b) require the same calculations; in (b) you have to interpret the results.

   (c) This is ±0.45 SE; the chance is about 35%.

   (d) 25%—use the binomial formula. (The sample is so small relative to the population that the chances are practically the same as if the draws were with replacement; hence the draws are nearly independent. The normal approximation cannot be used since the sample size n is so small.)
   $n = 10$, $k = 4$, $n - k = 6$, $p = 0.40$, $1 - p = 0.60$. The chance is then $\frac{10!}{4! \times 6!}(0.40)^4(0.60)^6 = 210(0.0256)(0.146656) = 0.25082 \approx 25\%$.

7. The average is

$$\frac{1+2+3+4=5+6+7}{7}=4;$$

the deviations are −3, −2, −1, 0, 1, 2, 3; the SD is

$$\sqrt{\frac{(-3)^2+(-2)^2+(-1)^2+0^2+1^2+2^2+3^2}{7}}=\sqrt{\frac{28}{7}}=\sqrt{4}=2.$$

(a) The standard error for the average is

$$\frac{\text{SD of the box}}{\sqrt{\text{number of draws}}}=\frac{2}{\sqrt{25}}=\frac{2}{5}=0.4.$$

Then 2 represents 5 times the standard error. So the chance that the result will be within 5 SE's of the EV is about 100%.

(b) Five of the seven tickets in the box are within 2 of 4. That represents 71% of the tickets in the box.

The typical ticket is off by 2, the SD, from the average of the box. That leads to the guess that 68% or so of the tickets will be off by 2 or less from 4.

However, the average of the draws will be off by an amount similar to the standard error of the average, which is 0.4, not 2.

(c) i. True.

ii. False. About 68% of the tickets should be within 1 SD, if this distribution is like the normal. And 1 SD is 2, not 0.4.

iii. False. You use the curve on the probability histogram for the average, not the histogram for the data.

iv. True. That's one standard error for the average, and with the number of draws equal to 25 and the initial box not too skewed this approximation should be pretty good.

8. Model: each measurement equals the exact elevation, plus a draw from the error box. The tickets in the box average out to 0. Their SD is unkown, but can be estimated by the SD of the data, as 30 inches. The SE for the sum of the 25 measurements is estimated as $\sqrt{25}\times 30=150$ inches. The SE for the average is estimated as $150/25=6$ inches.

(a) True.

(b) False: this mixes up SD and SE.

(c) False, same issue.

9. To decide this, a significance test is needed. The difference between the two percentages is 1.1%, and you need to put a standard error on this difference. Pretend that you have two independent random samples, drawn at random with replacement. The first sample is the prisoners who were assigned to the treatment group. There were 592 of them, and 48.3% of them were rearrested within a year of release. The SE for the number who were rearrested is

$$\sqrt{592} \times \sqrt{0.483 \times 0.517} \approx 12.16$$

The SE for the percentage rearrested is $12.16/592 \times 100\% \approx 2.0\%$.

The second sample is the prisoners who were assigned to the control group. There were 154 of them, and 49.4% of them were rearrested within a year of release. The SE for the number who were rearrested is

$$\sqrt{154} \times \sqrt{0.494 \times 0.506} \approx 6.2$$

The SE for the percentage rearrested is

$$6.2/154 \times 100\% \approx 4.0\%.$$

The SE for the difference is

$$\sqrt{2.0^2 + 4.0^2} \approx 4.5\%$$

On the null hypothesis, the difference between the percentages in the two samples is expected to be 0.0%. The observed difference is 1.1%. The test statistic is

$$z = \frac{\text{observed difference } - \text{expected difference}}{\text{SE for difference}} = \frac{1.1\% - 0.0\%}{4.5\%} = 0.24$$

P is about 40% $(\frac{100\% - 19.74\%}{2})$. This could easily be due to chance.

10. (a) $\chi^2 = 15.42$, 5 deg.fr., $P \approx 1\%$. (b) $\chi^2 = 6$, 5 deg.fr., $P \approx 30\%$.
(c) Pooled $\chi^2 \approx 15.42 + 6 = 21.42$, $5 + 5 = 10$ deg.fr., $P \approx 2.5\%$; conclude that the die is biased at the 5% level.

Calculations for part (a): all expected values are $600/6 = 100$.

$$\frac{(108-100)^2 + (93-100)^2 + (114-100)^2 + (120-100)^2 + (93-100)^2 + (72-100)^2}{100} = \frac{64+49+196+400+49+784}{100} = \frac{1542}{100} = 15.42.$$

Calculations for part (b): all expected values are $300/6 = 50$.

$$\frac{(57-50)^2 + (57-50)^2 + (51-50)^2 + (39-50)^2 + (54-50)^2 + (42-50)^2}{50} = \frac{49+49+1+121+16+64}{50} = \frac{300}{50} = 6.$$