review session solutions

Diagrams/pictures that would be necessary are not included in all solutions below!
NIB = Not in Book

p40, #18
(a) 115/192 or 60%
(d) 21/192 or 11%
(e) 35/77 or 45%
(f) no. if they were, then $35/50 = 36/44 = 06/21$.

p42 #27
(a) 34/365 or 9.3%
(b)90/365 or 24.7%
(c) $(27 + 268)/365$ or 81%
(NIB d) no. $27/7 \neq 63/268$. this means that the prediction of the weather person is not independent of the actual weather. by further inspection, we can conclude that the predictions were better than randomly guessing.

NIB 1
the distribution is skewed to the right with a median of *about* 150. the spread may be estimated by an IQR of *roughly* $300 - 75 = 225$. There are no unusual features (no outliers), and the distribution is unimodal.

NIB 2
(a) about the same
(b) the mean is greater than the median
(c) the mean is less than the median

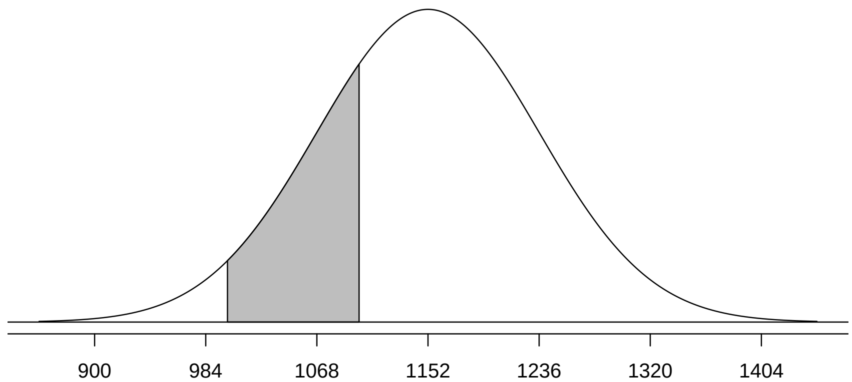
NIB 3
(a) 3; the median is the horizontal line inside the box.
(b) 3; Q_1, the first quartile, of 3 is about even with the median of 1 and 2. That is, the bulk of the data in 3 is larger than the bulk of the data in 1 or 2 (and significantly larger).
(c) 1
(d) 3

p135, #9
$\mu = 1152, \sigma = 84$
(a) $Z_a = \frac{x-\mu}{\sigma} = \frac{1000-1152}{84} = -1.81$. That is, 1000 pounds is 1.81 standard deviations below the mean.
(b) $Z_b = \frac{1250-1152}{84} = 1.17$. Since 1000 pounds is 1.81 standard deviations away from the mean and 1250 is only 1.17 s.d.'s away from the mean, 1000 pounds is more extreme.
(NIB c) Since this is a 90th percentile, Z will be positive. So, we can either look for 0.90 in the Z-table to get the positive Z or we can look for the 0.10 (the 10% tail), which will be a negative Z

(and we just make it positive). Either way, we should end up with

$$Z = 1.28 = \frac{x - \mu}{\sigma} = \frac{x - 1152}{84} \Rightarrow x = 1260 \text{ pounds}$$

(NIB d) $Z_d = \frac{1100-1152}{84} = -0.62$. This corresponds to a percentage of 27% (0.27 will be okay, too).
(NIB e) The area is shown below. This is the same as computing the area to the left of 1100 and subtracting off the area to the left of 1000: $0.268 - 0.035 = 0.233$ or 23.3%.



p362, #13

$P(\mathit{once}) = 0.17, P(\mathit{twice}) = 0.07, P(3+) = 0.04, P(\text{no repair}) = 1 - 0.17 - 0.07 - 0.04 = 0.72$
(a) 0.72
(b) $P(\text{no more than one}) = P(\text{no repair}) + P(\mathit{once}) = 0.72 + 0.17 = 0.89$
(c) $P(\mathit{some}) = 1 - P(\mathit{none}) = 1 - 0.72 = 0.28$
There are other ways to compute (b) and (c).

p362, #15

(a) $P(\text{neither need repair}) = P(\text{car 1 doesn't need a repair and car 2 doesn't need repair})$
$\overset{independence}{=} P(\text{car 1 no repair}) * P(\text{car 2 no repair}) = 0.72 * 0.72 = 0.52.$
(b) $P(\text{both need repairs}) = P(\text{1 needs repair and 2 needs repair})$
$= P(\text{1 needs repair}) * P(\text{2 needs repair})$
$= (1 - P(\text{1 doesn't need repair}))(1 - P(\text{2 doesn't need repair}))$
$= (1 - .72)(1 - .72) = 0.078.$
(c) $P(\text{at least one needs repair}) = 1 - P(\text{none need repairs}) = 1 - 0.52 = 0.48.$
(d) The way the cars are driven is not independent – the driver is the same. So, repairs are probably not the same either.

p386, #15

(a) P(no aces) = P(1. not an ace, 2. not an ace, 3. not an ace) = P(1. not an ace)*P(2. not an ace | 1. not an ace)*P(3. not an ace | 1. not an ace, 2. not an ace) = $\frac{48}{52}\frac{47}{51}\frac{46}{50} = 0.78$.
(b) P(1. red, 2. red, 3. red) = P(1. red)*P(2. red | 1. red)*P(3. red | 1. red, 2. red)
(26/52)(25/51)(24/50) = 0.12.
(c) P(at least one spade) = 1−P(no spades)= 1 − (39/52)(38/51)(37/50) = 0.59.
(NIB d) P(at least one king | 1, 2, and 3 are face cards) = 1−P(no kings | 1, 2, and 3 are face

cards)) = $1 - (8/12)(7/11)(6/10) = 0.75$

NIB 4

(a) P(6th dice is first 3) = P(1. not 3, 2. not 3, 3. not 3, 4. not 3, 5. not 3, 6. 3) = $(5/6)^5(1/6)$
(b) P(all even) = P(1. even, ..., 6. even) = $(3/6)^6 = 1/64$.
(c) P(4th dice is first 3) = P(1. not 3, 2. not 3, 3. not 3, 4. 3) = $(5/6)^3(1/6)$

p283, #9

Note: the decision isn't certain. We are basically looking at a random sample of 100 students.
(a) To simulate a vote, since there are 2 digits in the probabilities (0.45 and 0.55), there are 2 digits necessary in the simulation. For one voter, have the numbers 01-55 be a vote for 'my' candidate and 56-00 be a vote for the 'other' candidate.
(b) 20 elections, each with 100 voters is 2000 voters to simulate! a long time! (assuming we are doing this by hand and not using a computer program, which we have not discussed in this course.)
(c) 17: Y. 74: N. 06: Y. 26: Y. 40: Y. 88: N. 43: Y. 82: N. 64: N. 29: Y. In this simulation, my candidate got 6 of 10 votes.
(d) If we take a sample of size $n = 100$, then we can think of the proportion for our candidate as $\hat{p}$. So, we want to find P($\hat{p} < 0.50$). Note that conditions for normality are okay. So $\hat{p}$ is distributed normally with mean 0.55 and standard deviation $\sqrt{.55 * .45/100} = 0.05$. So, taking a Z-score for our cutoff of $x = 0.50$:

$$Z = \frac{.50 - .55}{.05} = -1$$

The probability of getting such a Z or a lower one, using the Z-table, is 0.16. So, the probability our friend loses the election, under the assumption that $p = 0.55$, is 0.16.
(NIB e) $H_0 : p = 0.55$. $H_A : p \neq 0.55$ (notice, we didn't believe it to be higher or lower prior to the vote!).

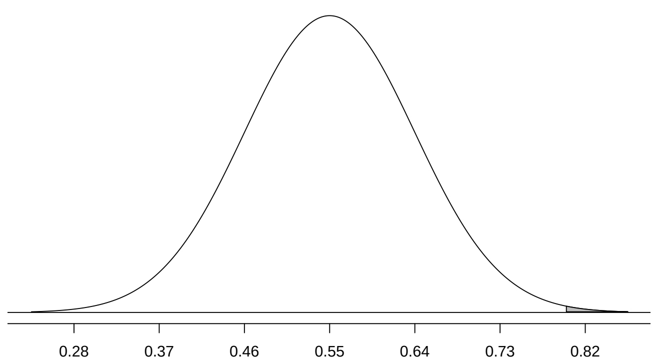


Figure 1: The area is over-represented in the plot – the area is too small to really be seen in a plot that is fully 'to scale'.

$Z = \frac{0.80-0.55}{\sqrt{.8*.2/100}} = 6.25$.
(NIB f) $SE = \sqrt{.8 * .2/100} = 0.04$ with $Z = 1.96$ ($\hat{p} = 0.80$). So the 95% CI is $0.80 \pm 1.96 * 0.04$, or 0.722 to 0.88.

p387, #21
(a) $P(pool|garage) = 0.266$.
(b) No. (One way to show this is $P(pool|garage) \neq P(pool)$, which it should if having a pool and garage are independent.)
(c) No. It is possible to have both.
(d) $P(neither) = 0.32$.

p388, #38
(a) If they were independent, then smoking would have no influence on whether there is a lung condition or not (but it does).
(b) P(lung condition)= 0.231.
(c, not in book) $P(smoker|condition) = 0.567$.

p425, #24
(a) a normal model would be appropriate (all assumptions are okay).
(b) P(number of cider apples $\leq$ 12) = P(proportion of cider apples $\leq$ 12/300) = 0.072. (Use normal model.)
(c) $3.64 * 10^{-15}$.
(d) Binomial would be most convenient. A normal model could also give a good approximation if setup appropriately. (Don't worry about this problem.)
(e) $H_0 : p = 0.06$. $H_A : p \neq 0.06$. We have already computed the tail probability in (c). The two-tail probability is then $7.28 * 10^{-15}$. Reject H_0 and conclude the true proportion is larger than 0.06.
(f) Here we use a slightly different standard error (computed using $\hat{p}$ instead of p). $(0.124, 0.209)$.

p457, #16
Question 1: The assumptions for a normal model are okay so it may be used. $SE = \sqrt{\frac{0.92*0.08}{160}}$. One tail probability only: 0.081.
Question 2: This probability would decrease if the number of seeds was 1600. (The $\hat{p}$ will give a better estimate of p, that is, be more 'tightly packed' around p when the sample size is larger.)

p542, #21
Run a two-sample Z-test (pooled). The alternative hypothesis should be one-sided. p-value is 0.0148. Reject H_0 and conclude the mammograms reduce the number of deaths from breast cancer.

NIB 6
We want the margin of error, ME, to be no more than 0.04. So we have $ME = Z * SE \leq 0.04$ and

$$1.96 * \sqrt{\frac{0.5(1-0.5)}{n}} \leq 0.04$$
$$n \geq \left(\frac{1.96}{0.04}\right)^2 * 0.25 = 600.25$$

Since n must be an integer and we want to ensure n is at least as large as 600.25, we choose n to be 601.

p604, #13

(b) American League: $(9.39, 10.23)$. National League: $(8.77, 10.39)$.

(f) $H_0 : \mu_{AL} - \mu_{NL} = 0$. $H_A : \mu_{AL} - \mu_{NL} > 0$. (Note that there was a prior belief that the AL teams would score more.) p-value of 0.30. Do not reject H_0. There is not sufficient evidence to say that the AL teams score more runs than the NL teams on average.

NIB 7

(a) Yes! One sample in the first survey corresponds to one sample in the second survey. Looking at the difference will be appropriate.

(b) No, the data is not paired. Each sample is independent of the other.

(c) The alternative hypothesis will be one-sided (thinking that the job satisfaction will be up in the second survey) for either test. The paired data gives a p-value of 0.000054, so reject the null hypothesis. The unpaired data (shown in the original description) yields a p-value of 0.325 and the we would not reject the null hypothesis.

Remark: The reason why the paired data in this case gives us more evidence is that we would think that the paired data gives us the **relative** increase in satisfaction of each individual. The standard deviation in the paired data was so small since it is often the case that a given individual's job satisfaction will change only a small amount (up or down), whereas one person's job satisfaction relative to another person's job satisfaction later in the month (the unpaired setup) would not be able to filter out the incremental increase of each individual. There are times when the second setup is better for detection but these conditions aren't straightforward and are beyond the scope of this course.

NIB 8

(a) (-0.0147,0.0016)

(b) Because 0 is in the interval, there does not appear to be a significant difference in femur length.

NIB 9

Not paired! (Not that there is really a way to do paired using the output anyways.)

$H_0 : \mu_{died} - \mu_{survived} = 0$

$H_A : \mu_{died} - \mu_{survived} \neq 0$

$t = -1.581$ (on 128.234 df), which corresponds to a p-value of 0.058. We would not reject the null hypothesis. We do not have sufficient evidence to say the femur length differs in birds that survived and those that died.

NIB 10

(a) $Humerus\_in = 0.172 + 0.785 * Femur\_in$.

(b) $\widehat{Humerus}\_in = 0.172 + 0.785 * (0.704) = 0.725$ inches.

(c) $residual = actual - predicted = 0.733 - 0.725 = 0.008$. The residual is positive, which means we underestimated the actual value.

(d) The absolute value of the correlation is given by $|r| = \sqrt{r^2} = 0.82$. That is, it is either -0.82 or $+0.82$. Recall that the slope of the line determines the sign of the correlation. Since the slope (given by 0.785) is positive so the correlation is positive: $r = 0.82$.

Remark: if you are uncertain of how to find the intercept, slope, and/or r^2 from the output, look in the R tutorial section.

(e) $r^2 = 0.67$ and it represents the amount of variation in $Humerus\_in$ explained by our model.