Intro to Prob and Stat I
GSIA, Carnegie Mellon University
45-733, Spring 2002 (mini 3)

Solutions

1. **In American roulette, there is a wheel with 38 spaces on it. The spaces are numbered 1-36, 0, and 00. There are eighteen red and eighteen black spaces, and the 0 and 00 spaces are green. At each play, each of the 38 spaces is equally likely to come up, and successive plays are independent.**

   There are many ways to bet roulette, but we will consider two ways. First, you may bet on either red or black. If the color you bet comes up, you win $1. If it does not, you lose $1. Second, you may bet on a particular number. If your number comes up, you win $35. If your number does not, you lose $1.

   (a) **(2 points) Betting black/red, what are the odds of winning?**
   There are 38 possible outcomes (the numbers 0,00,1-36), all of which are equally likely. Eighteen of these are red (and 18 are black), so if you bet on either red or black, your odds of winning are:
   $\frac{18}{38} = \boxed{0.474}$

   (b) **(2 points) Betting black/red, what is the expected value of playing?**
   If you win, you get $1, and if you lose you get -$1. These happen with probability 18/38 and 20/38, respectively, so the expected value of playing is:
   $\frac{18}{38}(1) + \frac{20}{38}(-1) = -\frac{2}{38} = \boxed{\text{-\$0.053}}$

   (c) **(2 points) Betting a specific number, what are the odds of winning?**
   There are 38 possible outcomes (the numbers 0,00,1-36), all of which are equally likely. One of these is the number you bet, so your odds of winning are:
   $\frac{1}{38} = \boxed{0.026}$

   (d) **(2 points) Betting a specific number, what is the expected value of playing?**

If you win, you get \$35, and if you lose you get -\$1. These happen with probability 1/38 and 37/38, respectively, so the expected value of playing is:

$\frac{1}{38}(35) + \frac{37}{38}(-1) = -\frac{2}{38} =$ \boxed{-\$0.053}

(e) **(6 points) In which way of betting is your chance of being ahead after 4 plays higher. (Ahead does not include "even," it means having more money than you started with)**

blk/red You are ahead after four plays if you have won three or four times. Let X be the number of wins in four plays. X has the binomial distribution. The probability of being ahead after four is $P\{X = 3 \bigcup X = 4\}$ or $P\{X = 3\} + P\{X = 4\}$.

$$\begin{aligned} P\{X = 3\} &= \frac{4!}{3!1!}\left(\frac{18}{38}\right)^3\left(\frac{20}{38}\right) \\ &= 0.224 \end{aligned}$$

$$\begin{aligned} P\{X = 4\} &= \frac{4!}{4!0!}\left(\frac{18}{38}\right)^4\left(\frac{20}{38}\right)^0 \\ &= 0.050 \end{aligned}$$

So, the probability of being ahead is $0.224 + 0.050 = 0.274$.

number You are ahead in this case unless you win 0 times. Let Y be the number of wins playing a specific number four times. The probability of winning zero times is:

$$\begin{aligned} P\{Y = 0\} &= \frac{0!}{0!4!}\left(\frac{1}{38}\right)^0\left(\frac{37}{38}\right)^4 \\ &= 0.899 \end{aligned}$$

So, the probability of being ahead after four plays is $1 - 0.899 = 0.101$.

So, the probability of being ahead after four plays is higher by playing \boxed{black/red}.

(f) **(6 points) If the casino runs a roulette game 1000 times a day, what is the probability that it comes out ahead that day? Assume only one person at a time plays and all bets are \$1 on a specific number.**

When the casino loses, it loses \$35, and when it wins, it wins \$1. To break even, it must, therefore, win $\frac{35}{36}$ plays, or $\frac{972.2}{1000}$ plays. To be ahead, it must win 973 or more plays.

So, let's calculate the probability that the casino wins at least a proportion 0.972 of the plays. (It would have been equally correct to use 0.973, arguably more correct)

Let X be the number of wins by the casino, and apply the CLT to the sample proportion to conclude:

$$\begin{aligned}
\frac{X}{1000} &\overset{A}{\sim} N\left(\frac{37}{38}, \frac{\frac{37}{38}\left(1-\frac{37}{38}\right)}{1000}\right) \\
\frac{X}{1000} &\overset{A}{\sim} N\left(0.974, 0.000025\right) \\
P\left\{\hat{p} > 0.972\right\} &= P\left(\frac{\hat{p}-0.974}{\sqrt{0.000025}} > \frac{0.972-0.974}{\sqrt{0.000025}}\right) \\
&\approx P\left\{Z > -0.4\right\} \\
&= P\left\{Z < 0.4\right\} \\
&= \boxed{0.6554}
\end{aligned}$$

2. **Your company frequently introduces new products. As part of this process, the marketing department is asked to forecast sales for each product's first year. You have been asked to evaluate the performance of marketing in this function. You have collected the following data (all in millions of $). You may assume that forecast and actual sales are distributed normally and independently.**

	Forecast	Actual
	25	16
	22	25
	15	8
	13	5
	36	29
	49	51
	18	17
	33	26
	22	19
mean	25.9	21.8
std dev	11.5	13.6

(a) **(5 points) Calculate a 90% confidence interval for actual mean sales. Interpret.**

Since the actual sales are distributed normally and since there are not enough observations to use a central limit theorem, we will use the t-table.

A 90% CI for E(Actual):

$$21.8 \pm 1.86\frac{13.6}{\sqrt{9}}$$
$$21.8 \pm 8.43$$

We are 90% confident that the true mean sales falls $\boxed{\text{between 13.4 and 30.2}}$.

(b) **(5 points) Calculate an 80% confidence interval for the actual variance of sales. Interpret.**

An 80% CI for V(Actual):

$$P\left\{a < \sigma_A^2 < b\right\} \quad = \quad P\left\{\frac{(n-1)s_A^2}{b} < \frac{(n-1s_A^2}{\sigma_A^2} < \frac{(n-1)s_A^2}{a}\right\}$$

$$
\begin{aligned}
&= P\left\{\frac{(n-1)s_A^2}{b} < \chi_8^2 < \frac{(n-1)s_A^2}{a}\right\} \\
&= 0.80
\end{aligned}
$$

Going to the chi-squared table:

$$
\begin{aligned}
\frac{(n-1)s_A^2}{b} &= 3.49 \\
b &= \frac{(n-1)s_A^2}{3.49} \\
b &= \frac{8(13.6)^2}{3.49} \\
b &= 424
\end{aligned}
$$

$$
\begin{aligned}
\frac{(n-1)s_A^2}{a} &= 13.36 \\
a &= \frac{(n-1)s_A^2}{13.36} \\
a &= \frac{8(13.6)^2}{13.36} \\
a &= 111
\end{aligned}
$$

We are 80% confident that the true actual variance lies between 111 and 424.

(c) **(5 points) Test, at the 10% level (2-sided) that mean actual sales are 29.**

$$
\begin{aligned}
\text{t} - \text{stat} &= \frac{\overline{X} - \mu}{\hat{\sigma}_{\overline{X}}} \\
&= \frac{21.9 - 29}{\frac{13.6}{\sqrt{9}}} \\
&= -1.59
\end{aligned}
$$

From the t-table, we get 1.86. Since $1.59 < 1.86$, we accept the null hypothesis and conclude that there is insufficient evidence in this sample to reject the hypothesis that the true mean of sales is 29.

There is a much easier way to answer this question. Notice that 29 is inside the 90% CI we calculated in part a. This means that this null will be accepted at the 10% or lesser level.

(d) **(5 points) If we were to test at the 5% level, what would happen?**

We would accept again. This is true since the 95% CI would be wider than the 90% CI (so that 29 would still be inside). Alternatively, since everything in the calculation in c would stay the same, except that the value from the table will increase, we would again accept.

3. Please use the same data as in the previous problem.

   (a) **(10 points) Calculate a 90% confidence interval for the mean error in forecast.**

   The error in forecast is just the forecast minus the actual value. So the mean error in forecast is $\mu_F - \mu_A$, the mean forecast minus the mean actual.

   The best approach to this problem is to recognize that this is a paired data problem. So, calculate each of the errors and then (treating the errors as a dataset) make a 90% CI for the mean.

   Doing this, you calculate a sample mean forecast error of 4.11 and a sample standard deviation for the forecast error of 4.51. Since we know the variables are normal and since we don't have enough observations to use a CLT, we use the t-table. The relevant value from the t-table is again 1.86.

   So, the 90% CI is:

$$4.11 \pm 1.86\frac{4.51}{\sqrt{9}}$$
$$4.11 \pm 2.80$$

   We are 90% confident that the mean error in forecast is $\boxed{\text{between 1.31 and 6.91}}$.

   There is an alternative method which is also correct. It is easy to calculate the sample mean forecast error from the information given in the problem, $25.9 - 21.8 = 4.1$. We can also calculate an estimate of the standard error if we use the fact (given in the problem) that forecast and actual sales are independent. This would be $\sqrt{\frac{11.5^2}{9} + \frac{13.6^2}{9}} = 5.94$. Then, the 90% CI would be $\boxed{4.1 \pm 11.0}$.

   (b) **(10 points) Test, at the 95% (typo! should be 5%) level, the hypothesis that the forecasts are right on average.**

   Using the calculations done above, we find that the $\text{t} - \text{stat} = \frac{4.1}{\frac{4.5}{\sqrt{9}}} = 2.73$ The t-table value for the 5% level is 2.305. The t-table does not have a value for 95%, so it would be acceptable to use the z-table's value of 0.065. Either way, we would $\boxed{\text{reject}}$ the null hypothesis that the forecasts are right on average — the marketing types consistently overestimate sales.

   If you calculated the variance in the alternative way above, you would get a $\text{t} - \text{stat} = \frac{4.1}{5.94} = 0.69$. This would lead you to accept the null if you were using the 5% level and to reject the null if you were using the 95% level.

4. **A health insurer is reviewing its contracts with hospitals. One important service it is considering is coronary artery bypass graft surgery ("heart bypass"). Death is an important complication of this surgery, so that death rates in hospital are an important quality indicator.**

   It is often claimed that "practice makes perfect" in this procedure, so you are to look into whether high volume providers (lots of practice) produce better outcomes. You compile the data available for your insureds and find:

Category	Patients	Dead
High Volume	1006	13
Low Volume	297	12

   (a) **(10 points) Test, at the 5% level, the claim that high and low volume hospitals have the same mortality rates. Interpret.**

   The null hypothesis is that $p_H - p_L = 0$ and the alternative hypothesis is that $p_H - p_L \neq 0$ (It would also have been OK to do a one-sided test here against $p_H - p_L < 0$). Because of the large number of observations, we can use a CLT to conclude that the sample proportions are distributed approximately normal and that we can use the z-table in our inference.

$$\begin{aligned} \text{t} - \text{stat} &= \frac{\hat{p}_H - \hat{p}_L - (p_H - p_L)}{\sqrt{\frac{\hat{p}_H(1-\hat{p}_H)}{n_H} + \frac{\hat{p}_L(1-\hat{p}_L)}{n_L}}} \\ &= \frac{\frac{13}{1006} - \frac{12}{297}}{\sqrt{\frac{\frac{13}{1006}(1-\frac{13}{1006})}{1006} + \frac{\frac{12}{297}(1-\frac{12}{297})}{297}}} \\ &= -2.296 \end{aligned} \tag{1}$$

   Since the z-table value is 1.96, we $\boxed{\text{reject}}$ and conclude that the high and low volume hospitals have different mortality.

   There is an alternative formula in the book for the standard error here. That formula is $\sqrt{\hat{p}_0(1-\hat{p}_0)\left(\frac{n_H+n_L}{n_H n_L}\right)}$, where $\hat{p}_0 = \frac{n_H\hat{p}_H + n_L\hat{p}_L}{n_H+n_L}$. If you used that formula, you got a t-stat equal to -3.05 and still rejected.

   (b) **(10 points) Compute an 80% confidence interval for the difference in mortality rates between high and low volume hospitals.**

   An 80% CI for $p_H - p_L$:

$$\begin{aligned} \hat{p}_H - \hat{p}_L &\pm 1.28\sqrt{\frac{\hat{p}_H(1-\hat{p}_H)}{n_H} + \frac{\hat{p}_L(1-\hat{p}_L)}{n_L}} \\ -0.0275 &\pm 0.0154 \end{aligned}$$

We are 80% confident that the true difference in mortality rates between high and low volume hospitals is between -0.0429 and -0.0121.

5. **As part of an effort to site a new plant, you perform a survey in Anytown, PA to assess local wage conditions. You survey, randomly, 100 workers in similar plants and find that they make, in wages and benefits, on average, \$23.12/hr with a standard deviation of \$6.75/hr.**

   (a) **(10 points) Make and interpret a 95% confidence interval for mean pay.**

   Notice that the sample size is pretty large, so we can apply a CLT and therefore use the normal table.

   A 95% CI:

$$\begin{aligned} 23.12 &\pm 1.96\frac{6.75}{\sqrt{100}} \\ 23.12 &\pm 1.32 \end{aligned}$$

   We are 95% confident that the true mean wage is between \$21.80/hr and \$24.44/hr.

   (b) **(10 points) Your boss wants a narrower interval. What are your options for giving it to her?**

   The two options I was looking for were:

   i. She could accept a lower confidence level. If she were to be willing only to be 80% confident (for example) that the correct answer is in the interval, I could provide a substantially narrower interval by using 1.28 in place of 1.96.

   ii. She could pay to collect a larger survey. This would raise n which would lower the standard error of the mean and make the 95% CI narrower

   I did accept one other possible option. Some students suggested narrowing the CI by lowering the standard deviation. However, to get credit for this answer, you would have had to make a concrete suggestion about how to lower it. The best of these suggestions was to define the job category more narrowly for the survey, so that the people responding would have a narrower range of pay.