Solutions to Statistics 571 Midterm 1

Hanlon/Larget, Fall 2010

1.

Solution:

(a) The sum of the probabilities must be one, so the missing probability is $1 - 0.15 - 0.25 - 0.2 - 0.3 = 0.1$. Let k be the missing value. We are given

$$\mathsf{E}(X) = 1.9 = (-5)(0.15) + (-1)(0.25) + (k)(0.2) + (4)(0.1) + (7)(0.3)$$

which simplifies to

$$1.9 = 1.5 + 0.2k$$

so $k = 0.4/0.2 = 2$.

(b) Use linearity of expectation.

$$\mathsf{E}(3X + 7) = 3\mathsf{E}(X) + 7 = 3(1.9) + 7 = 12.7$$

(c) One approach is to use the formula $\mathsf{Var}(X) = \mathsf{E}(X^2) - (\mathsf{E}(X))^2$.

$$\mathsf{E}(X^2) = (-5)^2(0.15) + (-1)^2(0.25) + (2)^2(0.2) + (4)^2(0.1) + (7)^2(0.3) = 21.1$$

So, $\mathsf{Var}(X) = 21.1 - 1.9^2 = 17.49$. As a check, this means the standard deviation is a little more than 4, and the points one standard deviation above and below the mean are about -2 and 6. Much of the probablity is within this range and some is outside; the value is consistent with our understanding.

(d) Use the properties of variance.

$$\mathsf{Var}(3X + 7) = \mathsf{Var}(3X) = 9\mathsf{Var}(X) = 157.41$$

2.

Solution:

(a) The distribution of individual fish lengths is $\mathrm{N}(54.0, 4.5^2)$. Let X be the length of a randomly chosen fish.

$$\mathsf{P}(X > 60) = \mathsf{P}\left(\frac{X - 54}{4.5} > \frac{60 - 54}{4.5}\right) = \mathsf{P}(Z > 1.33) \doteq 0.09176$$

using the table and a rounded Z score, or with R

```
> 1 - pnorm(60, 54, 4.5)
[1] 0.09121122
```

(b) The sampling distribution of $\bar{X}$ is normal with mean 54.0 mm and standard deviation $4.5/\sqrt{4} = 2.25$. Another way to say this is $\bar{X} \sim \mathrm{N}(54, \frac{4.5^2}{4})$.

(c)

$$\begin{aligned} \mathsf{P}(51 < \bar{X} < 60) &= \mathsf{P}\left(\frac{51 - 54}{2.25} < \frac{\bar{X} - 54}{2.25} < \frac{60 - 54}{2.25}\right) \\ &= \mathsf{P}(-1.33 < Z < 2.67) = 1 - 0.00379 - 0.09176 = 0.90445 \end{aligned}$$

using the table and rounded z scores. With R,

```
> pnorm(60, 54, 2.25) - pnorm(51, 54, 2.25)
[1] 0.9049584
```

(d) In the table, the z-score closest to the 0.9 quantile (right tail area equals 0.10000) is $z = 1.28$. The 0.9 quantile is then $54 + 1.28(2.25) = 56.88$. With R,

```
> qnorm(0.9, 54, 2.25)
[1] 56.88349
```

(e) The middle 80% is between the 0.1 and 0.9 quantiles. We already have the latter. The former is $54 - 1.28(2.25) = 51.12$. Again, with R,

```
> qnorm(c(0.1, 0.9), 54, 2.25)
[1] 51.11651 56.88349
```

(f) No. By the central limit theorem, the distribution of $\bar{X}$ is approximately normal even when the population is not normal when n is large enough, but $n = 4$ is not a large sample.

3.

Solution: A solution might include a tree to guide the calculations.

(a) Use the law of total probability to add the probabilities of both paths through the tree that end with color blind people.

$$(0.53)(0.02) + (0.47)(0.001) = 0.101107$$

(b) This is one path in the tree.

$$(0.47)(0.001) = 0.00047$$

(c) Use Bayes' Theorem.

$$\begin{aligned} \mathsf{P}(\text{man} \,|\, \text{color blind}) &= \frac{\mathsf{P}(\text{man} \cap \text{color blind})}{\mathsf{P}(\text{color blind})} \\ &= \frac{\mathsf{P}(\text{man})\mathsf{P}(\text{color blind} \,|\, \text{man})}{\mathsf{P}(\text{color blind})} \\ &= \frac{(0.53)(0.02)}{0.01107} \doteq 0.9575 \end{aligned}$$

4.

Solution:

(a) Of the 25 plants in the spider mite treatment group, 10 develop wilt disease. The observed proportion is $\hat{p} = 10/25 = 0.4$ but we will use the recommended method which adjusts the observed proportion. $p' = (10 + 2)/(25 + 4) \doteq 0.414$. The estimated standard error with the adjusted data is

$$\text{SE} = \sqrt{\frac{(0.414)(1 - 0.414)}{29}} \doteq 0.09146$$

and the margin of error is $1.96 \times 0.09146 = 0.179$. The 95% confidence interval is 0.414 ± 0.179 or $0.235 < p < 0.593$. In the context of the problem,

> We are 95% confident that the proportion of cottom plants given the spider mite treatment that will contract wilt disease after innoculation using the experimental conditions is between 0.235 and 0.593.

(b) It is sufficient to look at the graph and see that the curve reaches its highest point when $p = 0.40$, so $\hat{p} = 0.40$ and that the height of the curve at this point is between -1.8 and -1.9, closer to -1.8, say about -1.82.

Some of you figured out how to calculate this, although this was not necessary. As $n = 25$, if $\hat{p} = 0.4$ then $X = 10$. The log-likelihood is the natural logarithm of the binomial probability for 10 successes when $n = 25$ and $k = 10$:

$$\ell = \ln\left\{\binom{25}{10}(0.4)^{10}(0.6)^{15}\right\} \doteq -1.825$$

(c) The observed and expected counts are

Observed Counts

	Mites	No mites	Total
Wilt disease	10	20	30
No wilt disease	15	5	20
Total	25	25	50

Expected Counts

	Mites	No mites	Total
Wilt disease	15	15	30
No wilt disease	10	10	20
Total	25	25	50

where the upper left expected count is found by $(30)(25)/50 = 15$ and others are found similarly. The test statistic is

$$G = 2\big(10\ln(10/15)+15\ln(15/10)+20\ln(20/15)+5\ln(5/10)\big) \doteq 2(-4.055+6.082+5.754-3.466) \doteq 8.63$$

There are $(2-1)(2-1) = 1$ degrees of freedom, so the G statistic should be compared to the $\chi^2(1)$ distribution. From the table of quantiles, the p-value is $1 - 0.995 = 0.005$ when $G = 7.88$ and $1 - 0.999 = 0.001$ when $G = 10.83$. As the actual G is between these values, $0.001 <$ p-value < 0.005.

As an aside, as a $\chi^2(1)$ distribution is the same as a standard normal random variable squared, with a normal table we could have found the p-value as follows:

$$\mathsf{P}(X^2 > 8.63) = 2\mathsf{P}(Z > \sqrt{8.63}) \doteq 2\mathsf{P}(Z > 2.94) \doteq 0.0033$$

The factor of 2 comes from $\mathsf{P}(Z > \sqrt{8.63}) = \mathsf{P}(Z < -\sqrt{8.63})$ and either corresponds to $Z^2 > 8.63$. This trick only works when there is one degree of freedom.

The mean of a $\chi^2(1)$ random variable is 1, and $G = 8.63$ is quite a bit bigger. The small p-value is evidence against the null hypothesis of independence. In the context of the problem (and following the models from lecture notes),

> There is strong evidence (G = 8.63, $p < 0.005$, df=1, n=50, G-test for independence) that the spider mite treatment affects the probability of leaf wilt in the experimental conditions.

Soap box speech.— As biologists, we wish to summarize the strength of statistical evidence relevant to hypotheses of interest in order to persuade the informed and interested reader that our conclusions are supported. Keeping track whether or not a specific hypothesis is rejected or not at some arbitrary α level such as $\alpha = 0.05$ is unimportant. At the end of your academic career, there will be no one with a scorecard who tells you that, yes, you only failed to reject 5 percent of all of the true null hypotheses that you ever tested, and presents you with a gold star and a certificate of proper statistical decision making. People will care when they read your work if there is sufficient statistical evidence to back up the conclusions you draw *in the context of the data and your analysis for that single problem*. Do not think that drawing conclusions on the basis of a single number is sufficient for this purpose. Your conclusions will be based on the statistical analysis in part, but also other qualitative judments about the study design and background and information not directly included in the analysis. When summarizing results of a hypothesis test, it is better to report a p-value as a summary of the information from the hypothesis test than to make a statement whether or not the test is rejected.

5.

Solution:

(a) The population is Americans between the ages of 40 and 75 with osteoarthritis of the knee and severe pain when walking (50–90 on a 100 point scale).

(b) The sample is not random—all subjects are volunteers recruited from one of 46 study centers.

(c) Correct answers are **bold**.

Variable	**Explanatory** or **Response**	**Categorical** or **Quantitative**	**Experimental** or **Observational**
Treatment Group	**Explan.** or Resp.	**Cat.** or Quant.	**Exper.** or Obs.
Age	**Explan.** or Resp.	Cat. or **Quant.**	Exper. or **Obs.**
Dose of Tanezumab	**Explan.** or Resp.	Cat. or **Quant.**	**Exper.** or Obs.
Change in WOMAC pain subscore	Explan. or **Resp.**	Cat. or **Quant.**	Exper. or **Obs.**
Response to therapy	Explan. or **Resp.**	**Cat.** or Quant.	Exper. or **Obs.**

Notes:

- The first three variables are used to model changes in the last last two.
- Treatment group is the categorical variable version of quantitative variable dose.
- Age is used to define the population of interest, but it is observational because the researchers cannot assign a given subject to a specific age.
- As described in the footnote to the second table, response to therapy is a categorical designation based on a threshhold of the quantitative variable change in WOMAC pain score.