Guided Tour — Gelman & Hill, Chapter 2

1 Probability Distributions

This section presents an extremely compact review of basic facts about probability theory.

Normal Distribution: Means and Variances This section reviews some basic facts about the Normal distribution, the Central Limit Theorem, and Linear Combinations.

Some comments:

1. The bell-shaped normal curve arises frequently in nature because of the phenomena summarized in the *Central Limit Theorem*. Sums of independent random variables tend toward a normal distribution, so long as certain conditions hold.

2. Here are some additional facts worth mentioning:

   - The sum of a group of variables is a special case of a *linear combination*. The general behavior of linear combinations (in particular their mean, variance, covariance, correlation) is discussed in detail in the Psychology 310 lecture notes. A special case of these results is that, if the variables summed are *independent*, then the mean of their sum is the sum of their means, and the variance of their sum is the sum of their variances. However, it is important to realize that the theory underlying these results assumes that the linear weights defining the linear combination are constants, not random variables!
   - If you are not thoroughly familiar with the facts about linear combinations, you should go to the Psychology 310 page and review the lecture notes on linear combinations.
   - Sums of independent normal variables are always normal (as are all linear combinations of independent normal variables). Sums of *multivariate normal* random variables are always normal. However, sums of normal non-independent random variables need not be exactly normal.
   - G&H discuss the heights of all adults, and point out that the distribution of heights of U.S. adults is much less normal than the distribution of the heights of men (or women). There are several

ways of looking at these facts, and G&H choose one. However, another way of looking at this is as follows. The overall distribution of adult heights is a *mixture* of two distributions, men and women. In general, mixtures need not have the same shape as the distribution being mixed.

- On page 14, G&H point out that linear regression coefficients have an approximate normal distribution. Out of thin air, they give the formula for (sample) linear regression weights as $(X'X)^{-1}X'y$. This is a well-known result in multivariate statistics, when X contains the predictors and y the criterion. Note that each row of $(X'X)^{-1}X'$ takes a different linear combination of the scores in y. If y has a reasonably large number of scores in it, then the Central Limit Theorem Effect will take hold and indeed the resulting regression coefficients will have an approximately normal distribution. A technical point omitted in this discussion is that X in classic regression is considered fixed, not random. Strictly speaking, this model is inappropriate for many if not most applications of linear regression, but the assumption makes things much more mathemetically tractable. Notice that if X is fixed, so is $(X'X)^{-1}X'$, and so you are applying a constant set of linear weights to y.

3. On page 15, G&H repeat the discussion from the previous page! This time, they point out that X is assumed to be constant. It is important to realize that if X is not a constant, then β is not a fixed linear combination of y and different theory holds. As a simple example, suppose that y has a normal distribution and X is fixed. Then β is a fixed linear combination of independent normal variables and must be *exactly* normal. Now suppose that X contains random variables. In this case, the *beta* weights are no longer necessarily normal.

The discussion of the Poisson distribution gives 3 examples. Let's build (and solve) some questions surrounding these examples:

- If the cancer rate in your county is 4.52, what is the probability that there will be 7 or more cancer cases this year? To begin with, we need to recall that the Poisson cumulative probability is `ppois` and the density is `dpois`. The Poisson is a discrete distribution and can take on any non-negative integer value. The probability of 7 or more cases is one minus the probability of 6 or fewer cases, so we need

```
> 1 - ppois(6,4.52)
```

```
[1] 0.1715203
```

What is the probability that the number of cases will be between 3 and 6? We can use a neat trick in R to compute this directly.

```
> sum(dpois(3:6,4.52))
```

```
[1] 0.6571387
```

How did that work? Dissecting the interior part of the statement, we see that the `dpois` function was applied to the range of values 3:6, thereby producing a vector of results,

```
> dpois(3:6,4.52)
```

```
[1] 0.1675919 0.1893788 0.1711985 0.1289695
```

When we summed those values, we got the total probability.

- In the second example, hits on a website are modeled as Poisson(380). What is the probability that the number of hits will be between 350 and 400?

```
> sum(dpois(350:400,380))
```

```
[1] 0.7960098
```

Here's a topical question for you grad students who are TAing. Suppose that you have an office hour, and in your experience, on average 1.2 students come to an office hour. What is the probability that no students will come to an office hour?

2 Statistical Inference

In this section, G&H take pains to point out the distinction between sampling error and modeling error. The key point is made in connection with a simple linear regression model. Even if we had the entire population, there would generally still be an error of regression estimate, because models virtually never fit perfectly.

There is a brief discussion of standard errors and parameter estimates. The standard error of the sample mean is

$$\sigma_{\overline{X}} = \sqrt{\sigma/N}$$

The standard error of the sample proportion $\hat{p}$ is

$$\sigma_{\hat{p}} = \sqrt{p(1-p)/N}$$

. The above formulae are used frequently to construct simple statistical tests and confidence intervals. In practice, of course, we do not know either σ or p, so we substitute consistent estimators in the above formulas. When we do that, the resulting formulas are *estimated* standard errors. So, for example, a strict notation would say

$$\hat{\sigma}_{\hat{p}} = \sqrt{\hat{p}(1-\hat{p})/N} \tag{1}$$

They also point out in a footnote that in the case where the sample proportion $\hat{p}$ is 0, this leads to a standard error of zero, which precludes meaningful confidence intervals or hypothesis tests. They refer to a neat paper by Agresti and Coull (1998) that suggests a simple correction, i.e., when $\hat{p} = 0$, substitute $(y+1)/(N+2)$, where y is the number of "successes" and N the sample size.

In general, G&H take a refreshingly practical yet rigorous approach to fitting models, and they like to concentrate on things that consistently matter, rather than things that might matter. G&H do not, in general, maintain the distinction between "standard errors" (which are constants) and "estimated standard errors" (which are random variables). G&H are certainly not alone in this notational imprecision, and asymptotically, it makes no difference and in practical terms it usually matters only a bit. However, as any typical textbook chapter on Student's t emphasizes, there is a difference and at small samples it can matter.

3 Classical Confidence Intervals

The classical approach to confidence interval generation using asymptotic normal theory is covered in great detail in my Psychology 310 course handout, "A Unified Approach to Some Standard Statistical Tests". Essentially, if you have a linear combination of parameters, and you have for each parameter and estimator and an estimate of its standard error, you can immediately

construct formulas for confidence intervals. In particular, if the estimates are independent, and the linear combination you are interested in is

$$\sum_{i=1}^{J} \hat{\theta}_j \tag{2}$$

then the standard error is

$$\sqrt{\sum_{i=1}^{J} c_j^2 \hat{\sigma}^2_{\hat{\theta}_j}} \tag{3}$$

So, for example, suppose you are interested in the difference between two proportions, p_1 and p_2 for two separate populations. The difference $\hat{p}_1 - \hat{p}_2$ is a linear combination with linear weights $+1$ and -1. So, the standard error of $\hat{p}_1 - \hat{p}_2$

$$\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{N_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{N_2}} \tag{4}$$

G&H begin by giving some code for the classic simple confidence interval. These are of the form

$$estimator \pm Z^* \times SE \tag{5}$$

where SE is the standard error and $Z*$ is a percentage point from either the normal or t distribution. In general, the percentage points for a 95% confidence interval are the 2.5 and 97.5 percentile points. A 68% confidence interval uses Z^* of 1. G&H mix a variety of minor computational variations in these brief examples. I'll add comments to the R code. Here we compute the 95% confidence interval for a single mean for a simple data set of 5 observations

```
> y ← c(35,34,38,35,37)
> n ← length(y)
> # compute N
> estimate ← mean(y)
> se ← sd(y) / sqrt(n)
> int.95 ← estimate + qt(c(.025,.975),n-1) * se
> # notice how both confidence limits are computed simultaneously
> # by inputting a vector of percentage points!
> int.95
```

```
[1] 33.75974 37.84026
```

On page 20, G&H give an example of a confidence interval produced by statistical simulation. In this example, samples of 500 men and 500 women have yielded the result that the death penalty was supported by 75% of the men and 65% of the women. We would like to estimate the *ratio* of support among men to that of women. A point estimate is 75/ 65 = 1.15. To construct a confidence interval, they do the following:

1. Treat the sample proportions as if they were population proportions
2. Pretend that the sample proportion has an exactly normal distribution (not a bad approximation when $N = 500$
3. Take 10,000 simulations of $\hat{p}_1$ and $\hat{p}_2$. For each, compute the ratio $\hat{p}_1/\hat{p}_2$.
4. The 95% confidence interval has endpoints given by the 2.5th and 97.5th percentile of the 10,000 simulated ratios.

Here is the commented code:

```
> ## set up men
> n.men <- 500
> p.hat.men <- 0.75
> se.men <- sqrt(p.hat.men*(1-p.hat.men)/n.men)
> ## set up women
> n.women <- 500
> p.hat.women <- 0.65
> se.women <- sqrt(p.hat.women*(1-p.hat.women)/n.women)
> ## do 10,000 simulations
>
> n.sims<-10000
> ##rnorm creates normal random numbers
> simulations.of.p.hat.for.men <- rnorm(n.sims,p.hat.men,se.men)
> simulations.of.p.hat.for.women <- rnorm(n.sims,p.hat.women,se.women)
> ## we divide the two vectors of simulated sample proportions all at once
> simulations.of.ratio <- simulations.of.p.hat.for.men / simulations.of.p.hat.
> int.95 <- quantile (simulations.of.ratio, c(.025,.975))
> int.95


    2.5%    97.5%
1.062624 1.252238
```

4 Classical Hypothesis Testing

Confidence intervals can be used to test a null hypothesis that a parameter is a specified value. If the confidence interval excludes the specified value, we reject the null hypothesis.

In this section, G&H give an example of testing a distributional hypothesis by examining an assumption inherent in the choice of distribution. The Poisson distribution has a variance that is equal to its mean. One can therefore falsify the assumption that the data follow a Poisson distribution by examining the ratio of the variance to the mean. If the variance is significantly higher than the mean, then the data show *overdispersion*, while if the variance is significantly lower than the mean, the data show *underdispersion*.

In many textbook treatments of classical hypothesis testing in analysis of variance, substantial attention is paid to the issue of multiple testing and *familywise error rate*. The problem is that if you do many tests, and all the null hypotheses are actually true, the probability that *at least one* false positive will occur quickly rises toward 1.

G&H take the strong position on p. 22 that this isn't much concern to them. This is based on the view that null hypotheses of equality are almost invariably false in science.

5 Problems with Statistical Significance

In this section G&H point out some pitfalls with using "statistical significance" or a p-value as a scientific indicator. A key point is that differences in significance need not be significant. The classic example from Psychology 310 is this. Suppose you reject the hypothesis that $\mu_1 - \mu_2 = 0$ but do not reject the hypothesis that $\mu_3 - \mu_4 = 0$. This does *not* mean that there is a significant "difference of differences" (or interaction effect). For example, you might have barely rejected the first hypothesis and barely not rejected the second.

6 55,000 Residents Desperately Need Your Help

This is a nice example of statistical inference in practice. It draws on a number of principles dealt with in the earlier sections. Try to follow it closely and delineate where (if anywhere) you have trouble following the argument.