Read M&M Chapter 6. Section leaders will decide how much emphasis to place on M&M §6.3. Read M&M §8.2 (probably need to read M&M §8.1 as well). Confidence intervals and significance tests for a single normal. Power functions. Comparison of two normals. Comparison of two proportions via normal approximation with estimated variances.

Many of the ideas discussed in Chapters 6 and 8 of M&M became directly relevant during a recent series of court cases, in which I was an expert witness for the Defense. The Defense contended that for several years Hispanics had been underrepresented on juries. I testified in the context of a challenge to the process by which juries were selected in Connecticut.

The challenge extended over a period of more than two years. Part of the Defense case was based on results from a questionnaire, which was filled out by almost all persons presenting themselves for jury service at court houses in the Hartford-New Britain (HNB) judicial district. More substantial arguments involved a detailed analysis of records maintained by the agency responsible for administering the jury system. At each stage, statistical reasoning was central to the Defense arguments.

## 1. The juror questionnaires

For the questionnaire, prospective jurors were asked to check off one of five standard race categories and to check off an ethnicity category (Hispanic or not). For the first month, April 1996, the responses were as shown in the table.

| | Hispanic | Non-Hispanic | no response | total |
|---|---|---|---|---|
| 1 (= Black) | 3 | 152 | 1 | 156 |
| 2 (= White) | 21 | 1459 | 4 | 1484 |
| 3 (= AmerInd) | 4 | 7 | | 11 |
| 4 (= Asian) | 2 | 18 | | 20 |
| 5 (= Other) | 33 | 8 | | 41 |
| 1+2+3 | | 1 | | 1 |
| 1+3 | | 2 | | 2 |
| 1+5 | | 1 | | 1 |
| 2+3 | | 4 | | 4 |
| 2+4 | | 1 | | 1 |
| 2+5 | 1 | 4 | | 5 |
| no response | 16 | 2 | 3 | 21 |
| total | 80 | 1659 | 8 | 1747 |

Of the 1739 persons who answered the ethnicity question, 4.6% indicated they were Hispanic.

According to the case law, one way to assess the results is to compare the fraction of Hispanics from the questionnaires with 6.57%, the percentage of Hispanics in the over-18 population of the judicial district, according to the 1990 Census. Such a comparison would make sense if the State were claiming that the system was designed to represent the 1990 proportions. Actually, neither side was making such a claim, but case law is case law, so the comparison was made.

The first question was: Does the 4.6% support a contention that Hispanics are underrepresented?

## 2. Hypothesis tests

The Law is quite confused about how to compare two percentages, when the purpose is to check whether one percentage is much lower than it should be. A statistical method was used in a famous Supreme Court case (*Castaneda v. Partida*, 430 U.S. 482; footnote 17 on page 496), regarding a challenge to the composition of grand juries in Texas:

> If the jurors were drawn randomly from the general population, then the number of Mexican-Americans in the sample could be modeled by a binomial distribution. See Finkelstein, The Application of Statistical Decision Theory to the Jury Discrimination Cases, 80 Harv. L. Rev. 338, 353-356 (1966). See generally P. Hoel, Introduction to Mathematical Statistics 58-61, 79-86 (4th ed.1971); F. Mosteller, R. Rourke, & G. Thomas, Probability with Statistical Applications 130-146, 270-291 (2d ed. 1970). Given that 79.1% of the population is Mexican-American, the expected number of Mexican-Americans among the 870 persons summoned to serve as grand jurors over the 11-year period is approximately 688. The observed number is 339. Of course, in any given drawing some fluctuation from the expected number is predicted. The important point, however, is that the statistical model shows that the results of a random drawing are likely to fall in the vicinity of the expected value. See F. Mosteller, R. Rourke, & G. Thomas, supra, at 270-290. The measure of the predicted fluctuations from the expected value is the standard deviation, defined for the binomial distribution as the square root of the product of the total number in the sample (here 870) times the probability of selecting a Mexican-American (0.791) times the probability of selecting a non-Mexican-American (0.209). Id., at 213. Thus, in this case the standard deviation is approximately 12. As a general rule for such large samples, if the difference between the expected value and the observed number is greater than two or three standard deviations, then the hypothesis that the jury drawing was random would be suspect to a social scientist. The 11-year data here reflect a difference between the expected and observed number of Mexican-Americans of approximately 29 standard deviations. A detailed calculation reveals that the likelihood that such a substantial depature from the expected value would occur by chance is less than 1 in $10^{140}$.
>
> The data for the 2½-year period during which the State District Judge supervised the selection process similarly support the inference that the exclusion of Mexican-Americans did not occur by chance. Of 220 persons called to serve as grand jurors, only 100 were Mexican-Americans. The expected Mexican-American representation is approximately 174 and the standard deviation, as calculated from the binomial model, is approximately six. The discrepancy between the expected and observed values is more than 12 standard deviations. Again, a detailed calculation shows that the likelihood of drawing not more than 100 Mexican-Americans by chance is negligible, being less than 1 in $10^{25}$.

How would the method endorsed by the Supreme Court apply to the questionnaire data? Consider just the data in the last row of the table. We would start from the tentative assumption (the ***null hypothesis***) that the questionnaires represented a random sample from a population with 6.57% Hispanics. Compared to the size of the population, the sample was small. The removal of a small sample of persons from the population would have had a negligible effect on the percentage Hispanic in the rest of the population. It would hardly matter whether we treated the questionnaires as a simple random sample or as a result of sampling (with replacement) from a fixed population. Under the model, we could treat each questionnaire response like a toss of a coin that lands heads (Hispanic) with probability $p_0 = 0.0657$. The total number of heads $X$ from $n = 1739$ tosses would then have a $\text{Bin}(n, p_0)$ distribution. We could ask for the chance that $X$ would take a value of 80 or smaller. Using my computer, I calculated

$$\mathbb{P}\{X \leq 80\} = 0.0003 \qquad \text{if } X \text{ has a Bin}(1739, 0.0657) \text{ distribution.}$$

The 0.0003 is called a ***p-value***. Here p is short for probability; it does not refer to the $p_0$ for the Binomial distribution.

We are faced with a choice: either we regard the null hypothesis as reasonable, and write off the questionnaires as an occurence of a rare (a 3 in 10000 probability under the model); or we reject the null hypothesis.

If we did not have a computer that could calculate Binomial probabilities, we could use a normal approximation. Under the null hypothesis, the random variable $X$ has approximately a normal distribution, with mean = $\mu_X = np_0 = 114.3$ and standard deviation $\sigma_X = \sqrt{np_0(1-p_0)} = 10.3$.

REMARK. It is a good idea to choose one method for describing the outcome, then stick with it. For the cases in which I was involved, the results were usually described as percentages: for example, $4.6\% = 100 \times 80/1739\%$ of the potential jurors indicated they were Hispanic. It is unfortunate that so many different quantities are descibed as percentages. There is a great potential for confusion.

If a random variable $X$ has a $N(114.3, 10.3)$ distribution then

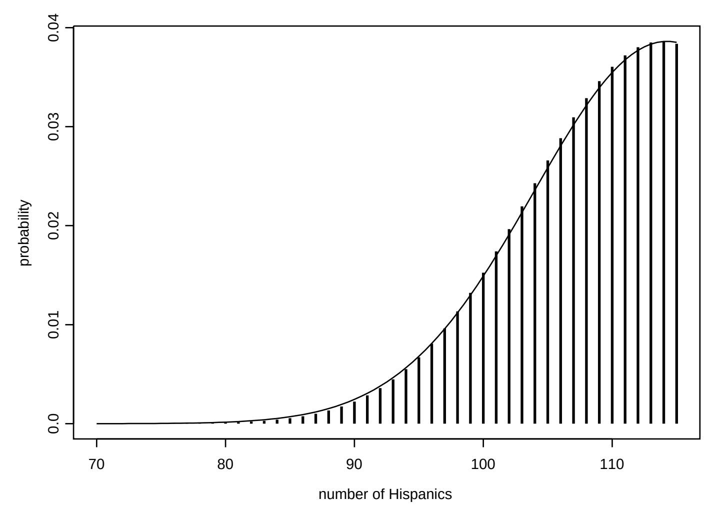
$$\mathbb{P}\{X \leq 80\} \approx 0.00046$$

The normal approximation gives a value reasonably close to the 0.0003 calculated for the Binomial.

Some authors prefer to treat hypothesis testing as a sharp accept-or-reject. They would fix a ***significance level***, such as 0.05, then reject the null hypothesis if and only if the p-value turned out to be smaller than the chosen significance level. If a random variable $X$ has a $N(114.3, 10.3)$ distribution then

$$\mathbb{P}\{X \leq 97.25\} \approx 0.05$$

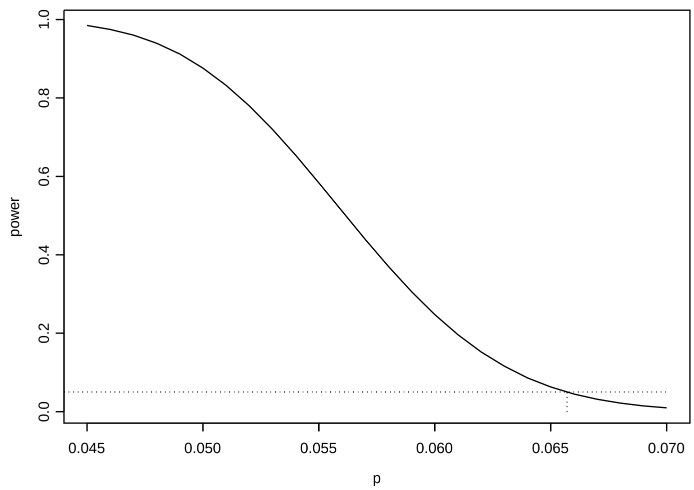
The test can be expressed as: reject the null hypothesis if and only if the number of Hispanics in the sample (the questionnaires) is $\leq$ 97. (Compare with $\mathbb{P}\{Bin(1739, 0.0657) \leq 97\} \approx 0.0498$, according to my computer.)




The Bin(1739,0.0657) probabilities, with the approximating normal density superimposed.

**Power of a test**

With a sharp accept-or-reject method of testing, it is possible to determine the behaviour of the test under alternative models. For example, we could consider other possible explanations for how the data were generated. For example, we could entertain the possibility that the true proportion of Hispanics in the population from which the questionnaires were taken equals some value of $p$ different from 0.0657. We could calculate $\mathbb{P}_p\{X \leq 97\}$ for various $p$, the subscript indicating use of the Bin$(n, p)$ for the probability calculation. This function of $p$ is called the ***power function*** for the test. It gives the probability that the null hypothesis will be rejected, under various models.




Power function based on the rejection region $\{X \le 97\}$ calculated at various Bin(1737, $p$) alternative distributions.

The horizontal dotted line corresponds to a power of 0.05. The threshhold 97 was chosen to make the power at $p = 0.0657$ close to 0.05.

Armed with the power function, we could back up the rejection of the null hypothesis by pointing to the fact that the event $\{X \le 97\}$ is much more likely under other plausible explanations for how the data were generated.

## 3. Confidence intervals

Neither side in the court case took the null hypothesis—that the questionnaires represented a sample from a population with 6.57% Hispanics—as a serious explanation for what might truly be happening. For one thing, the 1990 Census figures were most likely out of date. More importantly, everyone recognized that the persons turning up at the court houses had already been subjected to a number of filtering procedures that undoubtedly had had a large effect on the proportion of Hispanics presenting themselves as potential jurors. For example, noncitizens and persons with inadequate command of English were excused at an earlier stage of the summonsing process.

We could regard the observed 4.6% Hispanic from the questionnaires as an estimate of the percentage Hispanic in some hypothetical population with true fraction of Hispanics equal to an unknown value $p$. That is, we could regard $X$, the observed number of Hispanics, as a random variable with a Bin($n$, $p$) distribution, with $n = 1739$ and an unknown $p$. The observed fraction of Hispanics, $\widehat{p} = X/n$ would have an approximate $N(p, \sigma_p)$ distribution, where $\sigma_p^2 = p(1-p)/n$.

Write $\mathbb{P}_p$ to denote calculations under the Bin($n$, $p$) model, using the normal approximation. We know that

$$\mathbb{P}_p\{p - 2\sigma_p \le \widehat{p} \le p + 2\sigma_p\} \approx 0.95 \qquad \text{for each } p.$$

(I have rounded the value 1.96 from the normal tables up to 2. The whole thing is, after all, just an approximation.) That is,

$$\mathbb{P}_p\{ \text{ the random interval } [\widehat{p} - 2\sigma_p, \widehat{p} + 2\sigma_p] \text{ contains } p \} \approx 0.95$$

Of course, if we don't know $p$ then we don't know $\sigma_p$. But if we are prepared to guess that $p$ equals $\widehat{p}$ then we should be prepared to guess that $p(1-p)$ equals $\widehat{p}(1-\widehat{p})$. That is, we know that, under the Bin$(n,p)$ model, the random interval

$$\left[\widehat{p} - 2\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}, \widehat{p} + 2\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right]$$

has approximately a 0.95 probability of containing the true $p$. This random interval is called an (approximate) ***95% confidence interval for*** $p$.

If we put $\widehat{p} = 0.046$ we get $[0.036, 0.056]$ for the interval.

Now comes the tricky part. You find yourself in the situation of talking with an oracle (or at least a being who knows the true $p$) who promises to tell the truth with probability (approximately) 0.95. He also asserts that the $p$ in which you are interested lies in the interval $[0.036, 0.056]$. Is this one of the cases where he is lying, or not?

It is not correct to assert that the interval $[0.036, 0.056]$ contains the unknown $p$ with probability approximately equal to 0.95. Why not? Perhaps some analogies will help you to see the distinction.

<1> **Example.** You are about to generate an observation on a $N(0,1)$ distributed random variable $X$. You assert, correctly, that with probability 0.95 the value of $X$ you generate will lie in the interval $[-1.96, 1.96]$.

The generator generates, and out pops the value 0.72. Do you now say that 0.72 lies in the range $[-1.96, 1.96]$ with probability 0.95? No. You know the truth, and you know that the observed value is in the interval.

Now consider a second run of the same procedure. This time the generator generates $-2.89$. Do you assert that $-2.89$ lies in the interval $[-1.96, 1.96]$ with probability 0.95? Again, no. □

<2> **Example.** Consider a more complicated example. Suppose I am about to generate for you an observation on a $N(\mu,1)$ distributed random variable $X$. I know the value of $\mu$ but you don't. You assert, correctly, that, with probability 0.95, the value of $X$ will lie in the interval $[\mu-1.96, \mu+1.96]$. That is, with $\mathbb{P}_\mu$ probability 0.95 it will be true that

$$\mu - 1.96 \le X \le \mu + 1.96$$

Equivalently, the interval $[X-1.96, X+1.96]$ will have probability 0.95 of containing the unknown $\mu$. (The subscript $\mu$ on the $\mathbb{P}$ indicates that the calculation is made under the $N(\mu,1)$ model.)

Suppose the generator generates a value of $X$ equal to 17.35. You calculate $17.35 - 1.96 = 15.39$ and $17.35 + 1.96 = 19.31$. What can you say about the interval $[15.39, 19.31]$? Does it contain the true $\mu$? Can you now make any probabilistic assertion about whether it contains $\mu$ or not? Remember: I know $\mu$; I know whether $\mu$ lies in the interval or not. □

## 4. Difficulties in the interpretation of the questionnaires

The preceding discussion has avoided several crucial questions regarding the hypothetical population that generated the responses on the questionnaires.

One obvious difficulty is: How should we treat those persons who did not answer the question? Might not all those persons have been Hispanic? How could we counter an assertion that Hispanics have a strong propensity to refuse to answer questionnaires. I was able to point to other characteristics of the nonrepondents (surnames, and their answers to the race question, in some cases) to argue that there was no strong response bias.

The State raised more serious objections, by noting that:

(i) Persons summonsed to different court houses seemed to have different characteristics (such as different no-show rates). Moreover, some court houses had either opted out of the questionnaire collection altogether, or their administration of the questionnaires had been suspect. There was a court house effect.

(ii) Only persons who actually showed up at the court house filled in the questionnaire. Many potential jurors had their service cancelled the night before they were supposed to show up. There was no way of knowing whether someone whose service was cancelled would have turned up.

Objection (i) can be countered in part by focussing on responses from the two main court houses, to which most of the potential HNB jurors were summoned, and where the administration of the questionnaires was demonstrably quite careful and reliable.

Objection (ii) became a complication when the questionnaire results were to be compared with other administrative records. I had estimated the proportion of Hispanics amongst those potential jurors who were not disqualified for various reasons (such as inadequte command of English, or noncitizenship). The State argued that Hispanics were more likely not to turn up when summonsed, and that the percentage Hispanics on the questionnaires underestimated the percentage Hispanic in the population of those jurors who were not disqualified. The Defense did not disagree with this objection. Indeed, the Defense based its arguments on the estimates derived from the administrative records, and argued that the questionnaires were of use mainly as a consistency check.

The key question hidden behind these two objections is: What exactly should be inferred from the questionnaire responses?

## 5. Comparison of two proportions

At a later stage in the court proceedings, the Defense argued that the undeliverable rate rose during each court year, the period corresponding to summonses for September through August of the year.

For example, for the 1994-95 court year the numbers of summonses mailed and the numbers returned by the Post Office as undeliverable are shown in the next table.

| | Sep94 | Oct94 | Nov94 | Dec94 | Jan95 | Feb95 | Mar95 | Apr95 | May95 | Jun95 | Jul95 | Aug95 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| undeliverable | 775 | 776 | 917 | 925 | 963 | 1171 | 1502 | 951 | 1066 | 1130 | 991 | 1130 |
| total | 6393 | 6525 | 7449 | 7007 | 7149 | 8110 | 10532 | 6273 | 7028 | 6802 | 6061 | 7280 |

Model the number $X_1$ of undeliverable summonses for the first six months as an observation on a $\text{Bin}(n_1, p_1)$ distribution, where $n_1 = 6393 + \ldots + 8110$. Model the number $X_2$ of undeliverable summonses for the second six months months as $\text{Bin}(n_2, p_2)$ where $n_2 = 10532 + \ldots + 7280$. We would estimate $p_1$ by $\widehat{p}_1 = X_1/n_1$ and $p_2$ by $\widehat{p}_2 = X_2/n_2$. With the usual approximations, $\widehat{p}_1$ has approximately a $N(p_1, \sigma_1)$ distribution, where $\sigma_1^2 = p_1(1-p_1)/n_1$, and $\widehat{p}_2$ has approximately a $N(p_2, \sigma_2)$ distribution, where $\sigma_2^2 = p_2(1-p_2)/n_2$.

Assume independence between the two periods. Then

$$\text{var}(\widehat{p}_2 - \widehat{p}_1) = \text{var}(\widehat{p}_2) + \text{var}(\widehat{p}_1) = \sigma_1^2 + \sigma_2^2$$

The difference $\widehat{p}_2 - \widehat{p}_2$ has approximately a $N(p_2 - p_1, \sigma^2)$ distribution, where

$$\sigma^2 = \sigma_1^2 + \sigma_2^2 = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

REMARK. M&M page 601 seem merely to assert that a difference of two independent random variables, each normally distributed, is also normally distributed. This fact can be seen by writing the difference as a sum of a large number of independent pieces, then appealing for a central limit effect.

Under the null hypothesis that $p_1 = p_2$, the random variable $(\widehat{p}_2 - \widehat{p}_1)/\sigma$ has approximately a $N(0,1)$ distribution. We don't know $\sigma^2$, so we would have to estimate it. We might use

$$\widehat{\sigma}^2 = \frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}$$

or, as explained by M&M pages 604–605, we might argue that, under the model where $p_1 = p_2 = p$, the sum $X_1 + X_2$ would have a $\text{Bin}(n_1+n_2, p)$ distribution, so that we should estimate the common $p$ by $\widehat{p} = (X_1+X_2)/(n_1+n_2)$, and estimate $\sigma^2$ by

$$\widehat{\sigma}^2 = \frac{\widehat{p}(1-\widehat{p})}{n_1} + \frac{\widehat{p}(1-\widehat{p})}{n_2} = \widehat{p}(1-\widehat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

Whichever way you choose to estimate $\sigma^2$, you should end up with a random variable $(\widehat{p}_2 - \widehat{p}_1)/\widehat{\sigma}$ that should have an approximate $N(0,1)$ distribution if $p_1 = p_2$.

How would you you use such a statistics to test the hypothesis that the undeliverable rate was actually the same in both halves of the year, with alternatives where $p_2 > p_1$ being of the greatest interest?