Seth Zimmerman
Chetan Mehta
Math 50

Hot or Not: An Investigation of Attractiveness Distributions in Men and Women

I. Introduction

How do perceptions of male attractiveness differ from perceptions of female attractiveness? This paper addresses one aspect of that problem by attempting to determine whether and how the distribution of male attractiveness differs from the distribution of female attractiveness—i.e., if a man is selected at random, what are the chances he will fall within a certain attractiveness range, and how will those chances differ from the chances a woman will fall within that same range? Intuitively, several answers to this question seem plausible. On one hand, it seems anecdotally to be true that there are few extremely attractive people, many average looking people, and few extremely unattractive people. Such logic could lead one to predict a normal attractiveness distribution. On the other hand, one could also argue that there are two distinct prototypes for people—an "attractive" prototype and an "unattractive" prototype. This reasoning might lead to a prediction of bimodality in the attractiveness distribution. The work here aims to see whether empirical analysis bears out either of these cognitively plausible models, or whether it suggests that a different process may be at work in either the male or the female case.

Our investigation relies on a unique but somewhat compromised data source: the website www.hotornot.com. Users of this website rate pictures other users have posted of themselves on 1-10 attractiveness scale, with one being the low end and ten the high end. The website then computes and displays the average score each user has received, along with the total number of ratings the average is based on. With 75,000 pictures posted as of several years ago (http://www.hotornot.com/pages/about.html) and millions of votes cast in total, the site represents a potentially robust source of attractiveness data. It is clear, however, that the site's audience may not be very representative of the population as a whole—it is restricted to computer users, who may tend to be younger than the general population, and it is most likely frequented by people who are more concerned with their appearances than the average person. These caveats, however, do not prevent the hotornot.com data from providing an interesting and convenient first look at perceptions of attractiveness in men and women.

II. Data Collection

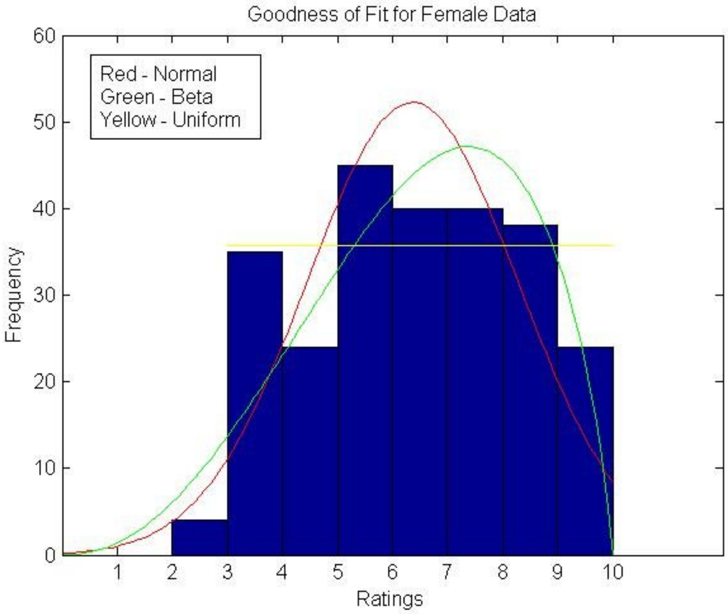
We collected a sample of 250 male scores and 250 female scores. Importantly, we ensured that each of these scores represented the average of at least 100 votes. This step was necessary to prevent the variance of the scores (which, as above, correspond to sample means of votes) from interfering with our plot of "true" attractiveness values, which we assumed the "true" averages would represent.

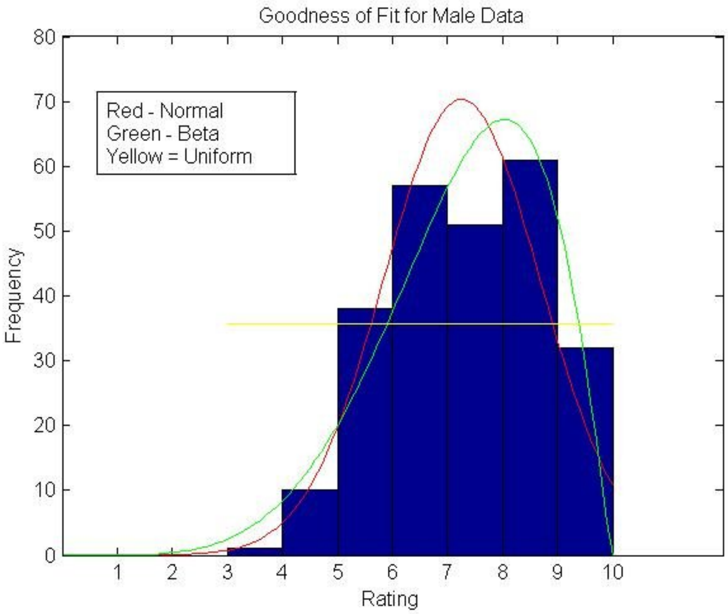
Establishing a 100 vote minimum allows for the construction of a 95% confidence interval of approximate width 1.75 even in an implausible worst case scenario. Since the variance of a score is equal to the variance of an individual vote, Y_i, divided by the total number of votes (i.e., $Var(\bar{Y}) = \sigma^2/n$, where σ^2 is the variance of an individual vote), the variance of the score for any person with at least 100 votes must be less than the variance of the score for a person with 100 votes, each of maximal variance. Clearly, the variance of an individual vote is maximized if it has a 50% chance of being a 1 and a 50% chance of being 10; i.e., for each vote Y_i, $P(Y_i=1) = .5$ and $P(Y_i=10)=.5$. Then, $E(Y_i)=5.5$, and $E(Y_i^2)=50.5$. So, $Var(Y_i)= E(Y_i^2)- E(Y_i)^2=50.5-30.25=20.25$. Assume we have taken 100 samples from a distribution with this worst-case variance. Then, $Var(\bar{Y})=20.25/100$, and we can set up a 95% confidence interval around $\bar{Y}$, given that $\sigma_{\bar{Y}} = \sqrt{Var(\bar{Y})}$ and

$2 * F_z(1.96) = .95$. Computing this interval yields a range around $\bar{Y}$ with approximate

(total) width 1.76 within which we can be 95% sure the real average value will fall. Since we do not intend to evaluate the data using bin widths of less than one unit in width, a 100 vote minimum seems acceptable to ensure that our average values will not diverge by more than one bin from the underlying true values—a precise enough fit to permit general distribution fitting, particularly given that it is almost impossible to imagine a case that would even approach the maximum variance.

III. Graphs, Data and Initial Observations





The relevant statistics for the data:

	Male	Female
Mean	7.2572	6.36
Variance	2.0083	3.6451
Standard Deviation	1.417	1.9092
N	250	250

Even without the above information, a glance at the histograms reveals differences in the data for men and women. The amount of data collected at the lower end of the distribution is small compared to the upper end. In addition, both histograms favor a bi-modal distribution – the female data more so than the male data – but since we had little experience with distributions of that nature, we were unable to analyze the data in that capacity.

IV. The Goodness of Fit Test

If attractiveness for men and women had some underlying distribution, we could test whether or not the data fit such a distribution. We decided to do Goodness-of-Fit tests, using procedures explained in Sections 10.3 and 10.4 of the textbook, which test whether or not it is plausible that a certain dataset comes from a defined distribution.

Our hypotheses were as follows:

H_0: $f_Y(y) = f_0(y)$

H_1: $f_Y(y) \neq f_0(y)$

The three different pdf's we tested were: normal pdf, uniform pdf, and beta pdf.

Normal: $f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$

Beta: $f(x;\alpha,\beta) = \frac{1}{\mathrm{B}(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}$

$\frac{1}{b-a} \quad \text{for } a < x < b$

Uniform: $0 \quad \text{for } x < a \text{ or } x > b$

The test statistic for the Goodness-of-Fit Test is:

$$D_1 = \sum_{i=1}^{t} \frac{(O_i - np_i)^2}{np_i}$$

D_1 is distributed with a X^2 distribution with $t\text{-}1\text{-}s$ degrees of freedom where t = # of outcomes and s = # of estimated parameters.

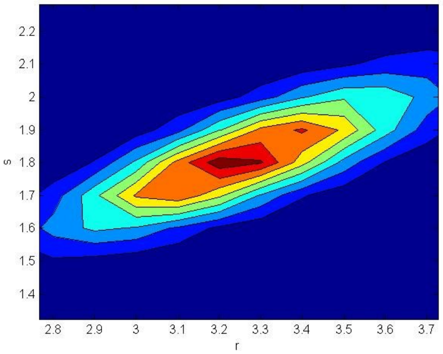
The goodness-of-fit procedure is summarized below:

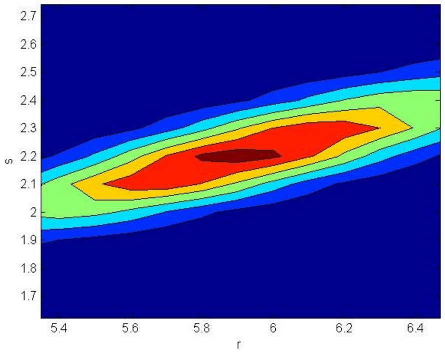
1. Divide ratings into *t* outcomes e.g. 0-1, 4-5, 9-10 etc.
2. Estimate the probability of getting a certain outcome if data came from the presumed distribution: $p_{i.}$
3. Multiply (2) by n to get expected frequency: $e_i = np_i$ (Note: $np_i \geq 5$ for all i)
4. Subtract observed from expected ($O_i - e_i$), square the difference, divide by ei and add together *t* outcomes to get a Chi-squared random variable.
5. If variable $D_1 \geq X^2_{1-\alpha,\, t-1-s}$ then H_0 should be rejected.

In our case, the parameters were unknown, so we needed to estimate parameters using Maximum Likelihood Estimation for the Beta and Normal distribution. As we learnt in class, the MLE estimate for the true mean and variance of a normal distribution is just $\bar{Y}$, the sample mean, and s^2, the sample variance. For the uniform distribution between 3-10, we used a value of $1/\theta = 1/(b\text{-}a) = 1/(10\text{-}3) = 1/7$.

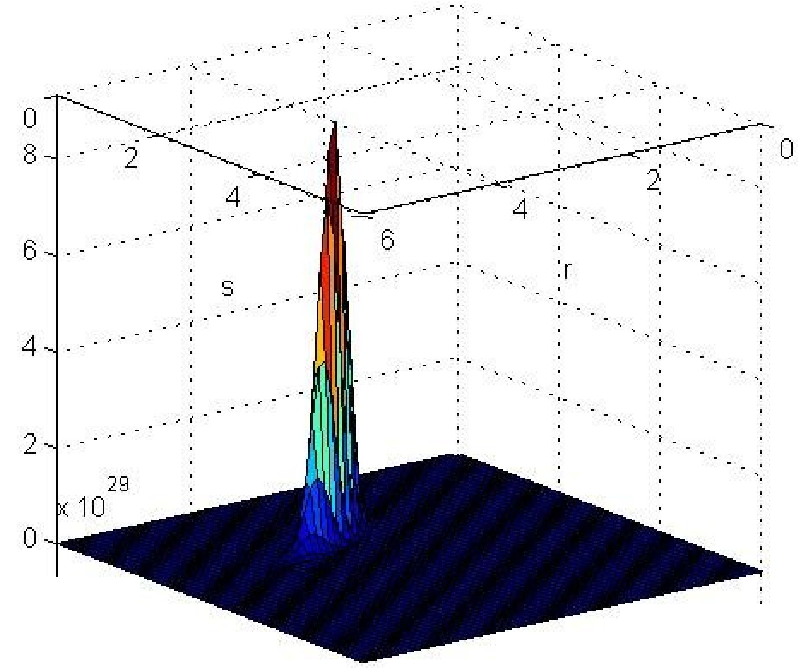
The Beta distribution was a little trickier; we were unable to devise an analytical method for obtaining the MLE's. We used the MATLAB command *betafit* to obtain the *r* and *s* values that best fit our data. Since we were unable to calculate MLE's rigorously, we did the next best thing: likelihood models for the parameters. Shown below are contour plots of a 2 parameter likelihood model for the beta pdf. As you can see, the red-hot areas correspond to the values obtained for *r* and *s* using the *betafit* command.

Female **Male**





We also have a 3D plot for the *r* and *s* values for the female data:



As can be seen, all three plots fit very well with our calculated parameters. We attempted to do a similar 3D plot for the male data, using a similar procedure, but ran into repeated problems with graphics rendering in MATLAB.

Thus, we fitted the following pdf's to the data:

	Normal PDF		Beta PDF		Uniform
	μ	σ	*r*	*s*	θ
Male	7.2572	1.417	5.9	2.20	7

Female	6.3600	1.909	3.25	1.8	7

V. Goodness of Fit Measurements

The tables below show our calculated values for the Goodness of Fit measurements. The 'Expectations' columns correspond to np_i and the 'Histogram' column corresponds to O_i. As you can see, in order to accommodate the $np_i \geq 5$ rule, we had to bunch together some of the bins in the tail of the distribution. This reduced the number of degrees of freedom and had a somewhat adverse impact on the resulting Chi-squared statistic.

Female							
Range	Histogram	Normal Expectations	Normal Errors	Beta Expectations	Beta Errors	Uniform Expectations	Uniform Errors
0-1	0	0.51616	0	0.44248	0	0	0
1-2	0	2.1748	0	3.4854	0	0	0
2-3	4	7.0046	3.3458	9.6701	6.7747	0	0
3-4	35	17.249	18.268	18.245	15.388	35.714	0.30229
4-5	24	32.48	2.214	28.005	0.57274	35.714	3.8423
5-6	45	46.772	0.067164	37.387	1.5501	35.714	2.4143
6-7	40	51.512	2.5728	44.447	0.44488	35.714	0.51429
7-8	40	43.389	0.26478	46.733	0.97	35.714	0.51429
8-9	38	27.951	3.6127	40.901	0.20573	35.714	0.14629
9-10	24	13.77	7.6007	20.685	0.53127	35.714	3.8423
			Sum = 37.945		Sum = 26.437		Sum = 11.576

The Chi-square values at the 0.05 significance levels and t-1-s degrees of freedom for the distributions and the results:

Normal: Critical value – 11.07, Observed – 37.945 => Reject the null hypothesis
Beta: Critical value – 11.07, Observed – 26.437 => Reject the null hypothesis
Uniform: Critical value – 12.95, Observed – 11.576 => Fail to reject the null hypothesis

Male							
Range	Histogram	Normal Expectations	Normal Errors	Beta Expectations	Beta Errors	Uniform Expectations	Uniform Errors
0-1	0	13.75	0.633	6.35	4.507	0	0
1-2	0					0	0
2-3	0					0	0
3-4	1					35.714	33.7
4-5	10			13.75	1.0227	35.714	18.5
5-6	38	33.375	0.6409	28.525	3.12	35.714	0.148
6-7	37	59.8	0.1311	47.4	1.9	35.714	12.708
7-8	51	67.47	4.02	63.3	2.39	35.714	6.557
8-9	61	48.05	3.49	62.55	0.0384	35.714	17.929
9-10	32	20.625	6.27	28.125	0.533	35.714	0.383

			Sum = 15.185		Sum = 13.511		Sum = 89.925

The Chi-square values at the 0.05 significance levels and t-1-s degrees of freedom for the distributions and the results:

Normal: Critical value – 7.815, Observed – 15.185 => Reject the null hypothesis
Beta: Critical value – 9.488, Observed – 13.511 => Reject the null hypothesis
Uniform: Critical value – 12.952, Observed – 89.925 =>Reject the null hypothesis

***Analysis*:** The only distribution that was not rejected was the Uniform distribution for Female data. This comes as somewhat of a surprise, since, upon visual inspection, it seems that the Beta distribution fits the Male data quite well. We hypothesize that additional data points, perhaps even 1 or 2 more in the lower tail-end, would've reduced the D_1 statistic substantially. The scarcity of data in that region produced very large errors for both male and female data, as can be seen in the tables. In fact, the single largest error for the Beta distribution when fitted to males comes in the first outcome, where there is a single data-point, but a larger number is expected. The squaring compounds the error and gives us a large test-statistic. This can be seen as an idiosyncrasy of the Chi-squared distribution.

VI. T-test on sample means

We conducted a T-test on the data in order to determine if the means were equal. The procedure is the same as the one used in class.

H_0: $\mu_f = \mu_m$

H_1: $\mu_f \neq \mu_m$

The critical value at the 95% level is 1.96. The combined sample variance was 2.8267 and the resulting T-value was 5.9663; thus we can safely reject the null hypothesis. The chances that the two means are equal are less than 5%. This does not come as much of a surprise considering the histograms.

VII. F-test on sample variances

We also conducted a F-test on the data to determine if the variances were equivalent. The procedure for the F-test was taken from Section 9.2 of the textbook. The rule for the 2-sided tests, which we conducted, is:

If S_y^2/S_x^2 is $\geq F_{a/2, m-1, n-1}$ or $\leq F_{1-a/2, m-1, n-1}$, then reject the null hypothesis.

$S_m^2 = 2.0083$
$S_f^2 = 3.6451$

m (not to be confused with male) = n = 250. So the test was conducted at 249 degrees of freedom in a 95% confidence interval.

Critical values: (0.813, 1.21)
Observed F-value = 2.0083/3.6451 = 0.5510

Therefore, we reject the null hypothesis; there is a <5% chance that the underlying variances are equivalent.

VIII. Applications

It is natural to ask at this point how the distributional differences discussed above affect the way we "experience" the data—what does a beta distribution, after all, have to do with real life? One way to understand these differences is to compare the probability of randomly picking a very attractive man from the HotorNot listings against the probability of randomly picking a very attractive woman from the HotorNot listings. If we define "very attractive" individuals as those who have a score greater than 7, we can construct a binomial distribution that reflects this probability. Consider women first: since 97 of the 250 women in the sample have scores higher than 7, we can estimate the true probability p that a woman on HotorNot is very attractive with the ML estimator $\hat{p} = {}^{97}\!/_{250}$. Further, to get an idea of how precisely this estimator reflects the underlying reality, we can construct a 95% confidence interval around it. Since $Var(\hat{p}) = \frac{(\hat{p}*(1-\hat{p}))}{n} = \frac{(.388*(1-.388))}{250}$, and, as above, $2 * F_z(1.96) = .95$, this confidence interval is $\left[\hat{p} + 1.96 \cdot \sqrt{Var(\hat{p})}, p - 1.96 \cdot \sqrt{Var(\hat{p})}\right] = [.3276, .4484]$. So, if you bump into a random woman who's listed on HotorNot, we can confidently say that the chances she will be very good looking are (roughly) somewhere between one third and four ninths. On the other side of the spectrum, if we define someone who is very unattractive as anyone who has a score below four, an analogous procedure reveals that the ML estimator for p=(the chance a woman listed on HotorNot has a score below 4) is $\hat{p} = {}^{39}\!/_{250}$ and that a 95% confidence interval for $\hat{p}$ is [.111, .201]; i.e., if you bump into another woman who's listed on HotorNot, the chances that she'll be very unattractive are likely somewhere between one ninth and two tenths.

Comparing these results to equivalent statistics for the male sample produces startling results. 141 out of the 250 men fit the above definition of "very attractive," allowing, as above, the construction of the ML estimator $\hat{p} = {}^{141}\!/_{250}$ and of the 95% confidence interval [.502,.624] around that estimator. That is, we can say with 95% certainty that between ½ and 5/8 of men listed on HotorNot are "very attractive." The result for very bad looking men is even more surprising—only one of the 250 men in our sample scored less than a four. With such a low frequency, it becomes inappropriate to plot confidence intervals using the normal curve. Considering instead a Poisson distribution with $\hat{\lambda} = \hat{p} \cdot n = (1/250) \cdot 250 = 1$, we can estimate that the probability there are 0 or 1 attractive men in a group of 250 is almost ¾, while the probability that there are more than 4 attractive men in a group of 250 is less than 0.004. There are several ways to interpret the disparity between these results and those for women. The most obvious is to conclude that men are simply more attractive than women. Having become somewhat familiar with the website during the data collection process, I would argue that, at the very least, the differences between men and women are not as stark as they appear to be statistically. Rather, I would speculate that generally higher ratings for men, and, more notably, the almost total absence of men at the low end of the spectrum, are due primarily to differences in the attitude of the raters. Those who grade men have lower standards, so to speak, or perhaps are able to accept a broader variety of appearances. Those who grade women, on the other hand, may have a narrower concept of what it means to be attractive. Of course, these ideas are in keeping with the social consensus of our time, and so we should be careful in our decision to accept them in the face of ambiguous data (cf. *The Mismeasure of Man*). Regardless of our interpretation of the sex-based asymmetries in our data, we can draw from these results a Lake Wobegone message—most of us, men or women, are more attractive than "average" (i.e., 5), and many more of us are very attractive than are very unattractive.

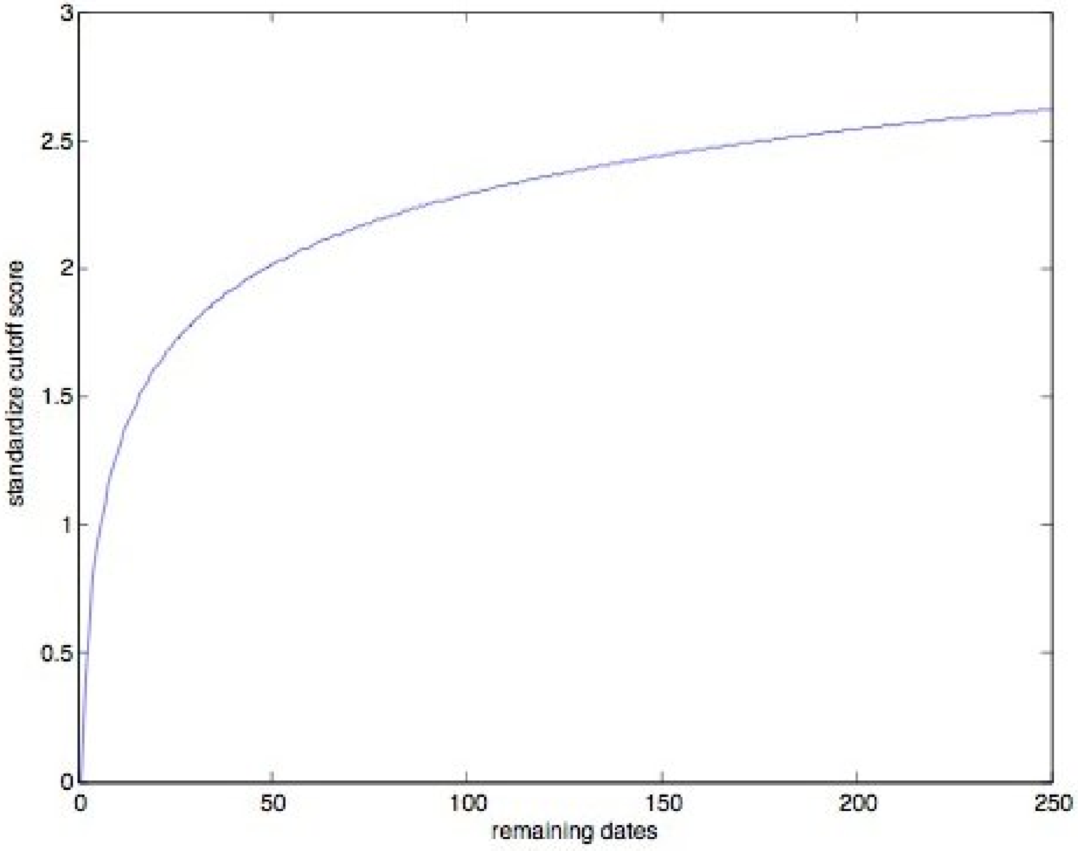
Another interesting problem involving this data begins with a very shallow character. Consider a man who wants to ensure that he chooses the prettiest out of *n* women from our sample. He is allowed to go on one date with each of them, and after that date, but before he sees any new women, he must choose whether to accept or reject that woman. If he accepts, he is committed to his choice, and if he rejects, he's lost his chance with that woman. How should this man choose who to accept and who to reject? For convenience, we'll assume here that the women are being pulled from a standardized normal distribution, that the man understands the distribution, and that he wants each of his decisions to maximize the expected score of the woman he ends up with.

First consider the case where *n*=2. That is, the man is a on a date, and knows he will have one more date after this one. Clearly, he should accept the current woman if she has a rating above the mean, his expected score if he tries his luck again, and reject her otherwise. Further, using standardized variables, the pdf of his score over two dates can be viewed in terms of split cases: $p_x(k) = \begin{cases} c \cdot \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, 0 \le x < \infty \\ \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, -\infty < x < \infty \end{cases}$, with the upper pdf being used if he accepts his first date and the lower pdf being used if he rejects his first date. Note that the upper pdf corresponds to the choice of a value from the top half of a normal distribution (with c being used as a normalizing constant), while the lower pdf is a standard normal.

Moreover, his expected score with this strategy over two dates is $(\frac{1}{2}) \cdot (\frac{2}{\sqrt{2\pi}}) \cdot \int_0^\infty xe^{-x^2/2}dx + (\frac{1}{2}) \cdot 0$, i.e., the probability he accepts his first date times his expected score given he accepts plus the probability he rejects his first date times his expected score given he rejects. Knowing this expected value makes it easy to compute his standards for the case *n*=3—if his third-to-last date has a score higher than his expected score over two dates, he should accept, and, if she has a lower score, he should reject. Generally, then, the critical score Y_n^*, above which he accepts and below which he rejects his *n*th-to-last date, should be determined by $Y_n^* = (1 - F_z(Y_{n-1}^*)) \cdot (\frac{c}{\sqrt{2\pi}}) \cdot \int_{Y_{n-1}^*}^{\infty} xe^{-x^2/2}dx + F_z(Y_{n-1}^*) \cdot (Y_{n-1}^*)$. That is, his critical score is equal to his expected value if he rejects his current date, which is equal to the probability he accepts his next date (i.e., the probability his next date has a score higher than Y_{n-1}^*, his cutoff score on that date) times his expected score if he accepts (where *c* is a normalizing constant chosen to set the area of the normal pdf between Y_{n-1}^* and ∞ to 1), plus the chances he rejects his next date times his expected score if he rejects (remember that (Y_{n-1}^*) was chosen because it is the expected score over *n*-2 dates). Note that *c* will always be equal to $\frac{1}{(1 - F_z(Y_{n-1}^*))}$ (since the area within the bounds of the integral is always $1 - F_z(Y_{n-1}^*)$), so the equation above simplifies to $Y_n^* = (\frac{1}{\sqrt{2\pi}}) \cdot \int_{Y_{n-1}^*}^{\infty} xe^{-x^2/2}dx + F_z(Y_{n-1}^*) \cdot (Y_{n-1}^*)$. Since computing the *n*th critical value requires only the (*n*-1)th critical value, a recursive algorithm is a natural computational choice here. Using such an algorithm, the standardized critical scores are as follows. The "W Score" column translates the critical score into the equivalent rating on the 1-10 scale using the sample mean and the ML variance estimator for the women's distribution, and the "M Score" column does the same conversion for the men's distribution.

Dates remaining	Critical Score	W Score	M Score
5	0.9127	8.1	8.6
4	0.7904	7.9	8.4
3	0.6297	7.6	8.1
2	0.399	7.1	7.8
1	0	6.4	7.3

The graph below shows how the standardized cutoff scores grow as the number of remaining dates increases. Note that it bears a close resemblance to a log function, with a steep increase in low n-values slackening quickly around the n=50 mark.



The moral of the story? If you're looking to maximize the expected score of your partner, it's important to know from the beginning that you're going to go on at least a few dates, but, after a certain point, it starts to matter less how many more you're planning on going on. Of course, this lesson depends entirely on the acceptance of the assumptions of normality and superficiality of judgment that we made when beginning this problem, so taking it seriously is not advisable.

IX. Conclusion

It would of course be a bad idea to draw any major conclusions about general attractiveness distributions from a relatively small sample of data from a non-representative source. Though our study does show that there are significant differences between the distribution of male and female ratings on HotOrNot, and that neither the male nor the female distribution conforms to our original hypotheses about the data, these discrepancies could be attributed to any number of factors, including differences between the men and women who choose to post their pictures, differences between the people who choose to rate men and the people who choose to rate women, and, possibly, differences between the social paradigms for male and female attractiveness. A more detailed, better-controlled study would be necessary to parse these variables.

Bibliography

http://www.hotornot.com. Accessed February 28 2006-March 7, 2006.

Gould, Stephen Jay. *The Mismeasure of Man*. New York: Norton and Company, 1981.