Stat 401, section 14.1 Goodness of Fit – Category Probabilities Specified

notes by Tim Pilachowski

Recall back to Lectures 6.1c, 8.4 (8.3 in the 8th edition) and 9.4 when we dealt with population proportions.

Vocabulary from 6.1c:

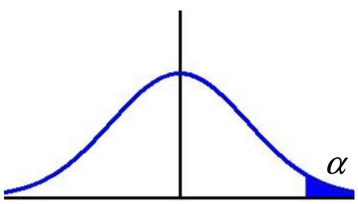
The point estimate for a population proportion was a sample proportion $\hat{p} = \dfrac{\text{number of successes}}{\text{total number tested}} = \dfrac{X}{n}$.

Hypothesis test for a single sample, from section 8.4 (8.3 in the 8th edition):
We did a little mathematical finagling with the the *z*-score formula for binomial distributions to get a *z*-score formula for proportions.

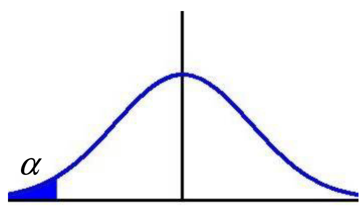
$$Z = \frac{X - np}{\sqrt{npq}} = \frac{\frac{X}{n} - \frac{np}{n}}{\frac{\sqrt{npq}}{n}} = \frac{\frac{X}{n} - \frac{np}{n}}{\sqrt{\frac{npq}{n^2}}} = \frac{\hat{p} - p}{\sqrt{pq/n}}$$

For the population proportion p in the numerator, and for the standard error (in the denominator), we used the value from the null hypothesis, p_0. Likewise, for the population parameter q, we used the null value $q_0 = 1 - p_0$. Note that we could express the numerator as "observed value minus expected value".

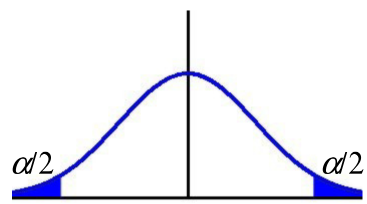
For the hypothesis test involving proportions, we used the same rule of thumb as we did for using a normal distribution to approximate a binomial distribution. If both $n * p_0 \geq 10$ and also $n * q_0 \geq 10$, then the hypothesis test is considered a large-sample test.



$H_0: p = p_0 \ (q = q_0)$
$H_a: p > p_0$



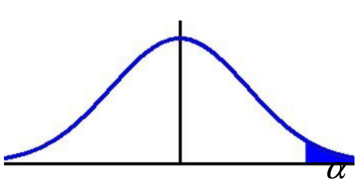
$H_0: p = p_0 \ (q = q_0)$
$H_a: p < p_0$



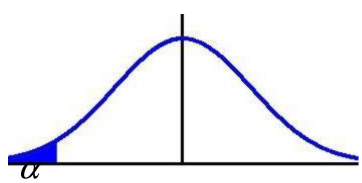
$H_0: p = p_0 \ (q = q_0)$
$H_a: p \neq p_0$

We tested to verify a large-sample situation, then computed the value of the test statistic: $z = \dfrac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}}$.

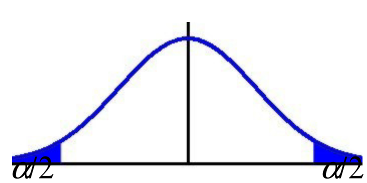
Hypothesis test for comparing two samples, from 9.4:
We only considered cases where $mp_1 \geq 10$, $mq_1 \geq 10$, $np_2 \geq 10$, $nq_2 \geq 10$, that is, a large sample case for which both sampling distributions, $\hat{p}_1 = \frac{X}{m}$ and $\hat{p}_2 = \frac{Y}{n}$, are approximately normal. Additionally, we focused solely on a hypothesis test for which $H_0: p_1 - p_2 = 0$ (which matches practice for the vast majority of actual situations).



$H_0: p_1 - p_2 = 0$
$H_a: p_1 - p_2 > 0$



$H_0: p_1 - p_2 = 0$
$H_a: p_1 - p_2 < 0$



$H_0: p_1 - p_2 = 0$
$H_a: p_1 - p_2 \neq 0$

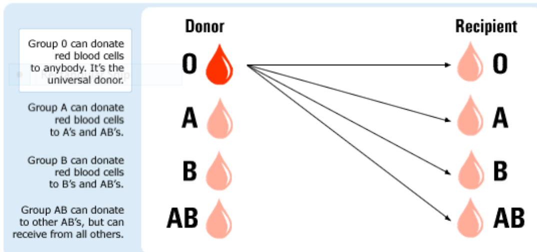
Verify a large-sample situation, and compute the value of the test statistic: $z = \dfrac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{m} + \frac{1}{n}\right)}}$.

Note that, once again, we could express the numerator as "observed value minus expected value".

In all of the previous encounters with population proportions, we were considering a **binomial experiment** consisting of a sequence of independent trials in which each trial could result in one of two possible outcomes: *S* (for success) and *F* (for failure). The probability of success, denoted by p, was assumed to be constant from trial to trial, and the number of trials, n, was fixed at the outset of the experiment.

A **multinomial experiment** generalizes a binomial experiment by allowing each trial to result in one of k possible outcomes, where $k > 2$.

Example A background. Human beings have the following blood types.



In addition to the A and B antigens, there is a third antigen called the Rh factor, which can be either present (+) or absent (–). In general, Rh negative blood is given to Rh-negative patients. Rh positive blood or Rh negative blood may be given to Rh positive patients.

So, there are eight categories of blood type ($k = 8$):

The universal red cell donor has blood type O−.
The universal plasma and platelet donor has blood type AB+.

Side note: While whole blood has a shelf life of 42 days, donated platelets must be used within 5 days of collection. There is a constant need for platelets, which are an essential part of the treatment for any low-platelet condition, including some types of cancer. You can donate platelets every seven days, up to 24 times a year.

According to the American Red Cross, proportions of the different blood types in the U.S. population are:

	Caucasian	**African American**	**Hispanic**	**Asian**
O+	37%	47%	53%	39%
O−	8%	4%	4%	1%
A+	33%	24%	29%	27%
A−	7%	2%	2%	0.5%
B+	9%	18%	9%	25%
B−	2%	1%	1%	0.4%
AB+	3%	4%	2%	7%
AB−	1%	0.3%	0.2%	0.1%

Why would this information be important? As one example, some genetic disorders are much more effectively treated by using blood from a donor who comes from the same ethnic group.

Notation: In general, we will refer to the k possible outcomes on any given trial as categories, and p_i will denote the probability that a trial results in category i. If the experiment consists of selecting n individuals or objects from a population and categorizing each one, then p_i is the proportion of the population falling in the i^{th} category. This type of experiment will be approximately multinomial provided that n is much smaller than the population size.

Why would we need n "much smaller than the population size"?

The null hypothesis of interest will specify the value of each p_i.

Example A. If we were to test the proportions of blood types among ethnic Hispanics in the U.S. population given in the table above, what would the null hypothesis be?

For a multinomial analysis, the alternative hypothesis will state that H_0 is not true – that is, that at least one of the p_i's has a value different from that asserted by H_0 (implying that at least two must be different, since all of the proportions added together must equal 1).
[This is similar to the statement of the alternate hypothesis in an ANOVA.]

Notation: The notation p_{i0}, "*p* sub *i* nought", will represent the value of p_i claimed by the null hypothesis.
[This is the same as the notation for the null value in linear regression analysis, β_{10}.]
In the "blood types among ethnic Hispanics" Example A above, $p_{10} = 0.53$, $p_{20} = 0.04$, $p_{30} = 0.29$, etc.

Theory: Before a multinomial experiment is performed, the number of trials that will result in category *i* ($i = 1, 2, \ldots, k$) is a random variable – just as the number of successes and the number of failures in a binomial experiment are random variables. This random variable will be denoted by N_i, and its observed value by n_i. Since each trial results in exactly one of the *k* categories, (that is, each observation in placed in exactly one category), it will always be true that $\sum_{i=1}^{k} N_i = n$, where *n* is the total number of trials. Likewise, the sum of the observed n_i's will necessarily be *n*. For example, in an experiment with $n = 50$ total trials and $k = 4$, random variable N_1 might take on value $n_1 = 10$, N_2 might take on value $n_2 = 15$, and N_3 might take on value $n_3 = 5$. Then N_4 must take on value $N_4 = 20$.

In other words, there is an underlying assumption that the *k* categories are *comprehensive*, and every observation will fit into one of those categories. The sum of the n_i's will equal the sample size *n*, and in terms of proportions, $\sum p_i = 1$.

For a multinomial analysis, the hypotheses will be stated in terms of *relative frequency*, p_i, but the test statistic will be calculated in terms of *frequency*, n_i.

In a binomial experiment, the expected number of successes and expected number of failures are *np* and *nq*, respectively. (Recall the formula for the mean of a binomial probability distribution, Lecture 3.4.) When $H_0: p = p_0$ $(q = q_0)$ is true, the expected numbers of successes and failures are $n * p_0$ and $n * q_0$, respectively.

Similarly, in a multinomial experiment the expected number of trials resulting in category *i* is

$$E(N_i) = np_i \quad (i = 1, 2, \ldots, k).$$

When $H_0: p_1 = p_{10},\ p_2 = p_{20},\ \ldots,\ p_k = p_{k0}$ is true, these expected values become

$$E(N_1) = np_{10},\ E(N_2) = np_{20},\ \ldots,\ E(N_k) = np_{k0}.$$

Example A revisited. A researcher wants to test the proportions of blood types among ethnic Asians in the U.S. population with $n = 800$.

What would the hypotheses be?

The expected frequencies when H_0 is true are

The n_i's and corresponding expected frequencies are often displayed in a tabular format. In this blood type Example, given Asian ethnicity, for n = 800 (and k = 8), this table might look like the one below.

Category	O+ (i = 1)	O– (i = 2)	A+ (i = 3)	A– (i = 4)	B+ (i = 5)	B– (i = 6)	AB+ (i = 7)	AB– (i = 8)	Row total
Observed	n_1	n_2	n_3	n_4	n_5	n_6	n_7	n_8	$n = 800$
Expected									

Note that the Expected values are those that are calculated assuming the null hypothesis is true.

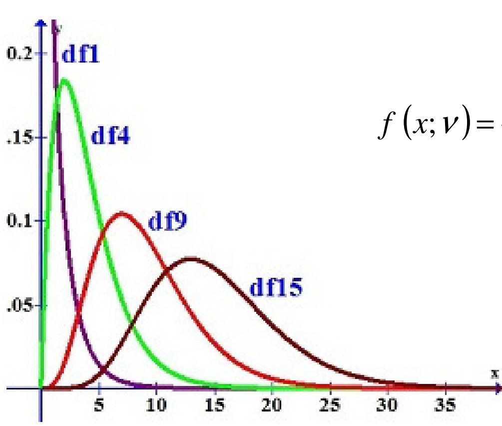
The Observed n_i's are usually referred to as *observed cell counts* (or *observed cell frequencies*), and Expected values $E(N_i) = n\,p_{i0}$ are called the corresponding *expected cell counts* under H_0. When H_0 is true, each n_i should be reasonably close to its corresponding $E(N_i) = n\,p_{i0}$. If, however, several of the observed counts differ substantially from their expected counts, we may have sufficient evidence to conclude that the actual values of the p_i's differ markedly from what the null hypothesis asserts.

The test procedure involves assessing the discrepancy between each Observed n_i and its associated Expected value $E(N_i) = n\,p_{i0}$, with H_0 being rejected when at least two discrepancies are sufficiently large.

We might, in a method similar to what we did when comparing multiple population means in Chapter 10, base our measure of discrepancy on the squared deviations $(n_1 - n\,p_{10})^2,\ (n_2 - n\,p_{20})^2,\ \ldots\ ,\ (n_k - n\,p_{k0})^2$, and calculate a sum of squares, $\sum (n_i - n\,p_{i0})^2$. However, since we are testing proportions, equal numeric differences might translate into very different proportions. In the text's example, the authors suppose $n_1 = 95,\ n\,p_{10} = 100$, $n_2 = 5$ and $n\,p_{20} = 10$. In both cases the squared numeric difference is 25. But, $n_1 = 95$ is only 5% less than its expected value $n\,p_{10} = 100$, while $n_2 = 5$ is 50% less than its expected value $n\,p_{20} = 10$.

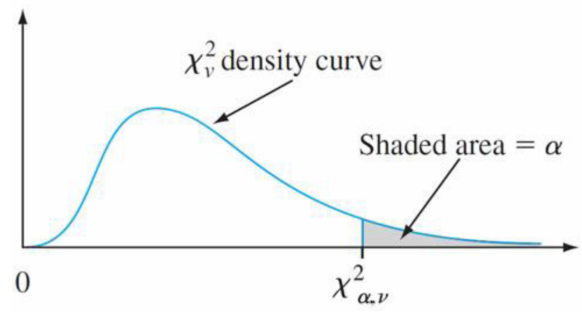
To take relative magnitudes of the deviations into account, we will take each squared deviation and divide it by its corresponding expected count: $\sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} = \sum \frac{(n_i - np_{i0})^2}{np_{i0}}$.
Note that this value will always be positive.

The probability distribution of this statistic is neither a Z nor a T distribution, but is rather one called a **chi-squared (χ^2) distribution**. (See text section 4.4. In my Stat 400 Lectures I bypassed this distribution.)



$$f(x;\nu) = \frac{1}{2^{\nu/2}\,\Gamma\!\left(\frac{\nu}{2}\right)} x^{(\nu/2-1)} e^{-x/2}, \quad x \ge 0$$

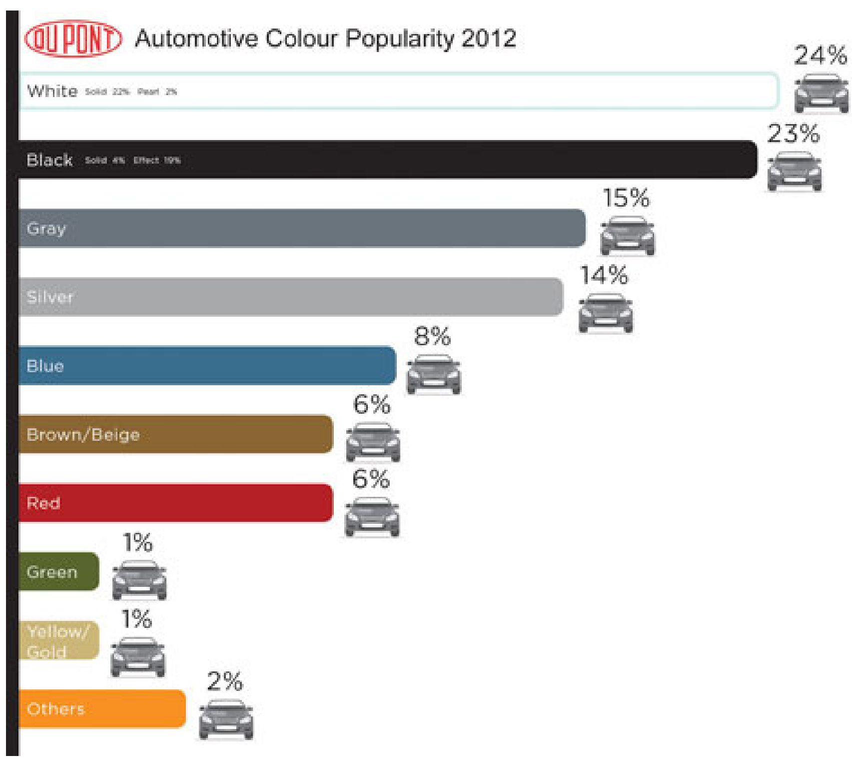
The chi-squared distribution has a single parameter, $\nu = k - 1$ degrees of freedom, with possible values 1, 2, 3, … . (There are only $k - 1$ "freely determined" cell counts: once any $k - 1$ are known, the remaining one is uniquely determined since the sum must equal n.) Analogous to the critical value $t_{\alpha,\nu}$ for the t distribution, $\chi^2_{\alpha,\nu}$ is the value such that α of the area under the χ^2 curve lies to the right of $\chi^2_{\alpha,\nu}$.



Selected values of $\chi^2_{\alpha,\nu}$ are given in Appendix Table A.7. We'll interpret the χ^2 test statistic in the same way that we interpreted the F statistic in Chapter 10. A value of the test statistic which is greater than the value of the critical value, calculated $\chi^2 \geq \chi^2_{\alpha,k-1}$, will imply P-value $\leq \alpha$, and we will reject the null hypothesis.

Just as with other hypothesis tests, we have some underlying assumptions. We must have a simple random sample of independent observations, for which there are $k > 2$ categories. All observations must be used. Data can be expressed either as frequency, or as relative frequency that is then converted to a frequency. The rule of thumb for being able to use a χ^2 test is that for all cells, we must have $np_{i0} \geq 5$ for all categories.

Example B: Each year, DuPont Automotive releases its Color Popularity Report, a study analyzing and predicting color trends throughout the world.



The new-car-sales manager of a dealership, knowing that color preferences can change from one year to the next, polls 200 recent customers and gets the following results: White, 23%; Black, 21%; Silver, 18%; Gray, 14%; Red, 8%; Blue, 6%; Brown/Beige, 6%; Green, 1%; Yellow/Gold, 1%; Others, 2%. Do the poll results indicate that the dealership should adjust the proportions of colors that they keep in their inventory ($\alpha = 0.05$)?

Category	Observed $n_i = n\,\hat{p}_i$	Expected $= n\,p_{i0}$	$(O-E)$ $= (n_i - n\,p_{i0})$	$(O-E)^2$ $= (n_i - n\,p_{i0})^2$	$\frac{(O-E)^2}{E} = \frac{(n_i - n\,p_{i0})^2}{n\,p_{i0}}$
White					
Black					
Gray					
Silver					
Blue					
Brown					
Red					
Green					
Yellow					
Other					
				$\chi^2 = \sum \frac{(O-E)^2}{E} = \sum \frac{(n_i - n\,p_{i0})^2}{n\,p_{i0}} =$	

IMPORTANT: The symbol χ^2 is a *notation*! Do not square the sum in the last column.

hypotheses:

calculations for Observed $n_i = n\,\hat{p}_i$:

calculations for Expected $= n\,p_{i0}$:

calculations for $(O-E)=(n_i - n p_{i0})$, $(O-E)^2=(n_i - n p_{i0})^2$, $\frac{(O-E)^2}{E}=\frac{(n_i - n p_{i0})^2}{n p_{i0}}$:

critical value:

conclusion:

We have applied the chi-squared test to a situation in which $k > 2$. However, it can also be used when $k = 2$. Not surprisingly, the the chi-squared test for $k = 2$ is mathematically equivalent to the "comparison of two population proportions test" we used in Lecture 8.4 (8.3 in the 8th edition). It can be shown that $(Z)^2 = \chi^2$ and $(z_{\alpha/2})^2 = \chi^2_{\alpha,1}$ so that $\chi^2 > \chi^2_{\alpha,1}$ if and only if $|Z| \geq z_{\alpha/2}$.

IMPORTANT: As is the case with all statistical test procedures, one must be careful not to confuse statistical significance with practical significance. A calculated χ^2 that is greater than a critical value $\chi^2_{\alpha,k-1}$ may be a result of a very large sample size rather than any practical differences between the hypothesized p_{i0}'s and true p_i's. Before rejecting H_0, the $\hat{p}_i$'s should be examined to see whether they suggest a model different from that of H_0 from a practical point of view.

Good news: We're not going to consider the use of Table A.11 to find *P*-values for the χ^2 test statistic. For this class, we'll either rely on software to calculate the *P*-value or we'll use the "rejection region" method used in Example B above. (You might recognize this as the same process used to evaluate the *F*-distribution test statistic in chapter 10.)

We'll also be skipping "χ^2 When the P_i's Are Functions of Other Parameters".

However, we *are* going to take a look at "χ^2 When the Underlying Distribution Is Continuous". The underlying concept is fairly straightforward. Let X denote the variable being sampled and suppose the hypothesized probability density function of X is $f_0(x)$. As in the construction of a frequency distribution in Chapter 1, subdivide the measurement scale of X into k intervals $[a_0, a_1)$, $[a_1, a_2)$, … , $[a_{k-1}, a_k)$. (Note that the left-side boundary is closed and the right-side boundary is open.)

The cell probabilities specified by H_0 are then $p_{i0} = P(a_{i-1} \le X < a_i) = \int_{a_{i-1}}^{a_i} f_0(x)\,dx$.

The cell intervals should be chosen so that $np_{i0} \ge 5$ for $i = 1, 2, \ldots, k$ to meet our rule-of-thumb criteria. In practice, the cells are often selected so that the p_{i0}'s, and therefore the np_{i0}'s, are equal.

We won't have time to do an Example, so I'll refer you to the text's Example 14.4.

The following notes might help:

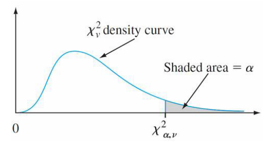
1) For the 90th percentile ("exactly 90% of all students will finish"), the critical value is z = 1.28.

2) The authors determined $\sigma = 15.63$ by replacing μ in "$\mu + 1.28\sigma = 120$" with its criterion value $\mu = 100$ and solving for σ.

3) The eight z-intervals were selected so that each has a probability of 1/8. That is, the probability is uniform for each interval, ensuring that the Expected value for each interval will be a uniform 120/8 = 15.

You should do the calculations for yourself for practice, checking your work against the text's results.

Appendix Table A.7
Critical Values for Chi-Squared Distributions



	α									
ν	**0.995**	**0.99**	**0.975**	**0.95**	**0.90**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**
1	0.000	0.000	0.001	0.004	0.016	2.706	3.843	5.025	6.637	7.882
2	0.010	0.020	0.051	0.103	0.211	4.605	5.992	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.344	12.837
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.832	15.085	16.748
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.440	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.012	18.474	20.276
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.534	20.090	21.954
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.022	21.665	23.587
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.724	26.755
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.041	19.812	22.362	24.735	27.687	29.817
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.577	32.799
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.407	7.564	8.682	10.085	24.769	27.587	30.190	33.408	35.716
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.843	7.632	8.906	10.117	11.651	27.203	30.143	32.852	36.190	38.580
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.033	8.897	10.283	11.591	13.240	29.615	32.670	35.478	38.930	41.399
22	8.643	9.542	10.982	12.338	14.042	30.813	33.924	36.781	40.289	42.796
23	9.260	10.195	11.688	13.090	14.848	32.007	35.172	38.075	41.637	44.179
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.558
25	10.519	11.523	13.120	14.611	16.473	34.381	37.652	40.646	44.313	46.925
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.807	12.878	14.573	16.151	18.114	36.741	40.113	43.194	46.962	49.642
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.120	14.256	16.147	17.708	19.768	39.087	42.557	45.772	49.586	52.333
30	13.787	14.954	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
31	14.457	15.655	17.538	19.280	21.433	41.422	44.985	48.231	52.190	55.000
32	15.134	16.362	18.291	20.072	22.271	42.585	46.194	49.480	53.486	56.328
33	15.814	17.073	19.046	20.866	23.110	43.745	47.400	50.724	54.774	57.646
34	16.501	17.789	19.806	21.664	23.952	44.903	48.602	51.966	56.061	58.964
35	17.191	18.508	20.569	22.465	24.796	46.059	49.802	53.203	57.340	60.272
36	17.887	19.233	21.336	23.269	25.643	47.212	50.998	54.437	58.619	61.581
37	18.584	19.960	22.105	24.075	26.492	48.363	52.192	55.667	59.891	62.880
38	19.289	20.691	22.878	24.884	27.343	49.513	53.384	56.896	61.162	64.181
39	19.994	21.425	23.654	25.695	28.196	50.660	54.572	58.119	62.426	65.473
40	20.707	22.164	24.433	26.509	29.050	51.805	55.758	59.342	63.691	66.766