Каждый опыт приобретает цену и значение только при использовании пригодных вспомогательных средств и при целесообразном их применении.

Грегор Мендель.

Критерий Колмогорова и экспериментальная проверка законов наследственности Менделя.

Ю. М. Барабашева, Г. Н. Девяткова, В. Н. Тутубалин, Е. Г.Угер.

По воспоминаниям одного из авторов данной статьи, лет сорок назад Андрей Николаевич Колмогоров примерно следующим образом пошутил на семинаре Лаборатории статистических методов. Он предложил участникам задуматься над таким вопросом. Известно ведь, что Франция имеет многовековую блестящую математическую школу. Почему же тогда во Франции применения математической статистики развиты гораздо меньше, чем в северных странах (т.е. в Англии, Дании, Швеции и т. д.)? Ответ А.Н.Колмогорова состоял в следующем: дело в том, что математическая статистика (как и закон) – что дышло: куда повернул, то и вышло. Чтобы это дышло поворачивать на пользу, а не во вред, нужно, кроме математической культуры, иметь известную моральную добропорядочность, которая у северных народов есть… (заключение о том, что у французов ее нет, потонуло в хохоте аудитории).

Сам Колмогоров оставил блестящий пример обращения с этим дышлом. Речь идёт о работе «Об одном новом подтверждении законов Менделя» [1]. Этот пример особенно ценен в следующем отношении. При изложении теории вероятностей и математической статистики крайне важно привести такие примеры, когда те или иные экспериментальные данные без применения вероятностных методов трактовались одним образом, а применение вероятностного подхода сделало трактовку данных иной, явно более правильной. Однако подобные примеры должны относиться к общепонятным научным вопросам, либо к вопросам, которые можно быстро объяснить (как, например, менделевские законы наследственности). Поэтому запас таких примеров, который можно извлечь из истории науки, ограничен, а работа [1] является одним из самых ярких примеров такого рода.

Научные связи А. Н. Колмогорова с рядом выдающихся биологов и обстоятельства появления работы [1] в настоящее время широко отражены в литературе (см., например, [3], [4], [5]). Однако, при ближайшем рассмотрении выяснилось, что материалы, связанные со статьей [1], проанализированы недостаточно детально. Возникла необходимость новой обработки этих данных, которая и представлена в предлагаемой статье.

Кроме того, методика А.Н. Колмогорова может быть применена для обработки классических результатов самого Г. Менделя и предполагает более детальную обработку данных, чем применение критерия хи-квадрат, сделанное, в частности, в работе Р.А. Фишера [6]. Представляет интерес сравнение выводов, получаемых при разных способах обработки, которое также излагается в данной статье. Рассматриваемая нами задача не выходит за рамки простых вероятностных моделей и ее можно использовать в процессе обучения для понимания того, как с помощью вероятностно-статистических моделей можно пытаться понять биологические явления.

1. Статистическое предисловие.

Что делает математическую статистику «дышлом», которое можно поворачивать по своему произволу (на самом деле, все же в определенных пределах)? Причина лежит в некоторой неопределенности самой той логики рассуждений, которая только и является возможной, если в результаты экспериментов вмешивается какая-то случайность.

1.1. Логика проверки гипотез.

Рассмотрим, в частности, статистическую проверку гипотез. Пусть в результате эксперимента возникает некий вектор $x=(x_1,x_2,\ldots,x_n)$, являющийся результатом n наблюдений, причем относительно вероятностей, связанных с вектором x, имеется некоторая гипотеза H. Как проверить по результатам наблюдений, верна или неверна эта гипотеза? На практике такая проверка осуществима лишь с той или иной потерей информации о наблюдениях $x_1,x_2,\ldots,x_n$. А именно: вводится некоторая функция $T=T(x)$, называемая *статистикой критерия*, которая в каком-то смысле оценивает различие между наблюдением x и неким теоретическим идеалом (определяемым гипотезой H). Непременное условие (для выполнения которого в конкретных случаях применена масса математической изобретательности) состоит в следующем: *распределение статистики* T (при верной гипотезе H) *известно*. Кроме того, подразумевается, что при мыслимых нарушения гипотезы H значения T имеют тенденцию увеличиваться. Таким образом, если гипотеза H верна, то значение $T=T(x)$, полученное для конкретных данных $x=(x_1,x_2,\ldots,x_n)$ не должно быть слишком большим. Количественно это оформляется одним из двух способов, сходных между собой, но все же несколько различных. При первом способе указывается некий порог t_α такой, что $P\{T\geq t_\alpha\}\leq\alpha$, и говорят, что гипотеза H отвергается на уровне значимости α, если в эксперименте получилось $T(x)\geq t_\alpha$. Этот способ называется проверкой гипотезы H *на уровне значимости* α.

При втором способе вычисляется p-значение, т.е. $p=P\{T\geq T(x)\}$, где $T(x)$ полученное в эксперименте значение. Иными словами, p-значение – это наименьший уровень значимости, на котором можно было бы отвергнуть гипотезу H, получив значение $T(x)$. Еще можно сказать, что p-значение – это вероятность получить *такое же или еще худшее* согласие с гипотезой H, какое получено в эксперименте. Понятно, что близкое к нулю p-значение ставит гипотезу H под сомнение. Выбор же порога для p-значения, ниже которого гипотеза H отвергается (равно как и выбор уровня значимости α) дает возможность некоторого произвола.

Но основной источник произвола – это выбор самой статистики критерия T. Уровни значимости, либо p-значения, отвечающие одной статистике T_1, могут не иметь ничего общего с соответствующими числами, отвечающими статистике T_2. При этом строгих критериев, определяющих выбор статистики, на практике чаще всего не существует. Это и есть то «дышло», которое, как и закон, можно поворачивать в любую сторону: по одному критерию T_1 будет получаться, что гипотеза H неверна, а по другому T_2 – что значение $T_2(x)$ лежит вполне в пределах допустимого.

Чем же ограничивается подобный произвол статистика? Надо сказать, что при однократной проверке гипотезы – ничем. Более того, выбор статистики T из ряда известных (относящихся к данному конкретному случаю) нередко производится после ознакомления с полученными в эксперименте данными $x=(x_1,x_2,\ldots,x_n)$. В этом случае уровни значимости и p-значения могут потерять последние остатки объективности. Однако ситуация меняется, если проверка гипотезы производится несколько раз на независимом экспериментальном материале. В частности, здесь помогает следующий математический факт: полученные в различных опытах p-значения $p_1,p_2,\ldots,p_k$ должны образовывать выборку из равномерного распределения на отрезке [0,1] (в том случае, когда гипотеза H верна и распределение статистики T непрерывно). Если гипотеза H неверна, то p-значения имеют тенденцию уменьшаться, т.е. будут накапливаться к левому концу отрезка [0,1].

Но что можно сказать в том случае, если p-значения, наоборот, накапливаются к правому концу отрезка [0,1]? Это ведь означает, что значения статистики T слишком малы, т.е. различие между наблюдениями и теоретическим идеалом меньше, чем полагалось бы при верной гипотезе H. В этом случае возникает подозрение, что экспериментатор фальсифицировал свои результаты с целью предъявить очень хорошее согласие с гипотезой. Именно этот последний ход мысли является тем стержнем, вокруг которого вращается обсуждение результата ряда генетических экспериментов, в том числе экспериментов самого Г.Менделя.

1.2.Критерий Колмогорова.

Критерий Колмогорова относится к тому случаю, когда числа $x_1,x_2,\ldots,x_n$, возникающие в эксперименте, представляют собой выборку. Это означает, что гипотеза H состоит в том, что эти числа суть значения n независимых и одинаково распределенных величин с функцией распределения $F(x)$. Критерий основан на статистике

$$T(x) = D_n = \sup_x |F_n(x) - F(x)|,$$

где $F_n(x)=\{$*число* $x_i<x\}/n$ –эмпирическая функция распределения (тех случайных величин, про которые говорит гипотеза *H*).

Как важнейшее свойство статистики D_n обычно подается то обстоятельство, что ее распределение вероятностей не зависит от функции $F(x)$ (при единственном условии, что эта функция непрерывна). До появления критерия Колмогорова в статистике рассматривались функции распределения вида $F(x;a_1,\ldots,a_k)$, определяемые конечным числом параметров $a_1,\ldots,a_k$. Оказалось, что от наличия конечного числа параметров можно отказаться (от этого становится только проще), и в этом смысле приемы, основанные на статистике D_n (и ряде аналогичных) называются *непараметрической статистикой*. Ее возникновение относят обычно к 1933 году, когда появилась работа [7], но в ныне опубликованных материалах имеется более раннее упоминание о занятиях Колмогорова этой тематикой. Именно в письме А.Н.Колмогорова П.С.Александрову от 1 мая 1931 г. (см.[8]) сказано следующее:

Кроме того, нашел одну прекрасную новую формулу из теории вероятностей:

Пусть F(x) закон распределения (непрерывный, но произвольный) и $F_n(x)$ *ступенчатое к нему приближение, найденное по n наблюдениям, тогда*

$$P\{max(F_n - F) \geq y\} \sim e^{-2ny^2}$$

Формула <u>совсем не зависит от характера закона</u> F(x), впрочем, формула еще не совсем доказана.

Таким образом, самооценка А.Н.Колмогорова также выдвигает на первый план независимость распределения статистики от теоретического закона $F(x)$.

Что касается статьи [7], то в ней изучены законы распределения статистики D_n при конечном n, а также указано предельное распределение статистики $\lambda = \sqrt{n}D_n$ [*] при $n\to\infty$. Иными словами, введена ныне знаменитая функция Колмогорова

$$K(y) = \lim_{n\to\infty} P\{\sqrt{n}D_n < y\}.$$

Приводится также небольшая таблица этой функции.

В настоящее время известны также формулы и таблицы, с помощью которых можно находить точное распределение статистики λ. Уже при n порядка нескольких десятков это точное распределение практически не отличается от $K(y)$.

Хотелось бы отметить, однако, что независимость функции распределения статистики D_n от теоретического закона $F(x)$ и способы ее вычисления – это, с точки зрения приложений, вопросы математической техники. На наш взгляд, гораздо более важно то обстоятельство, что применение критерия Колмогорова требует построения и визуального анализа эмпирической функции распределения $F_n(x)$. Эта функция суммирует в наглядном виде всю информацию, содержащуюся в наблюдениях $x_1,x_2,\ldots,x_n$, если считать, что порядок, в котором производились наблюдения, не важен. Вывод о том, годится ли некоторая функция $F(x)$ в качестве теоретического закона для наблюдений, базируется, прежде всего, на визуальном сравнении графиков функций $F_n(x)$ и $F(x)$. Этим более полным учетом информации критерий Колмогорова выгодно отличается от других статистических критериев, в частности от критерия Пирсона χ^2, при применении которого исходная информация учитывается менее полно.

2. Изложение содержания работы А.Н.Колмогорова [1].

Когда мы говорим о «моральной добропорядочности» статистика, упомянутой во введении, всегда возникает вопрос о точности действия статистической модели, которую разумно считать подтверждением (или опровержением) того или иного теоретического взгляда на вещи (в данном случае речь идет о менделевских законах наследственности). Подход А.Н.Колмогорова к этому вопросу заслуживает самого пристального внимания.

2.1. Общая постановка задачи.

Работа [1] начинается с постановки вопроса о том, чем отличаются точки зрения, с одной стороны, менделевской и моргановской генетики и, с другой стороны, школы Т.Д.Лысенко. Интересно, что при этом в уста Т.Д. Лысенко вкладывается вполне разумная и не противоречащая, в принципе, данным науки концепция. Согласно А.Н.Колмогорову, оба направления принимают, что свойства потомка двух родителей α и β определяются свойствами тех двух гамет, слияние которых в процессе оплодотворения было началом жизни этого потомка. Каждый из родителей производит очень большое число гамет,

[*] Статистикой Колмогорова обычно называют величину D_n (для малых n есть таблицы критических точек – см., например, [9]). Для больших n рассматривают статистику λ. Ее распределение определяется функцией Колмогорова $K(y)$, поэтому ее тоже иногда называют статистикой Колмогорова.

соответственно, $\alpha_1,...,\alpha_{k_1}$, и $\beta_1,...,\beta_{k_2}$, которые различаются между собой некоторыми биологическими особенностями. Лишь некоторые пары $(\alpha_{i_1},\beta_{j_2}),...,(\alpha_{i_n},\beta_{j_n})$ этих гамет дают потомство. При этом сторонники менделизма исходят из предположения, что вероятность выбора гамет, дающих потомство, не зависит от биологических особенностей этих гамет, так что все способы их выбора равновероятны (кратко это называется «гипотезой независимости»). Как и всякая другая гипотеза о независимости одних явлений от других, эта гипотеза может нарушаться (например, в случае неравной жизнеспособности гамет, либо селективного оплодотворения, либо неравной жизнеспособности потомства и т.д.).

А.Н.Колмогоров следующим образом формулирует различие между двумя точками зрения:

Серьезный спор может идти между такими двумя точками зрения.

1. *Гипотеза независимости в большинстве случаев является хорошим первым приближением к действительному положению вещей (сторонники менделевской и морган овской генетики).*[*]
2. *Селективное оплодотворение и неравная жизнеспособность играют всюду столь решающую роль, что рассмотрения, опирающиеся на гипотезу независимости, для биологии бесплодны (школа академика Т.Д.Лысенко).*

Как частный случай гипотезы независимости Колмогоров формулирует менделевский закон расщепления признаков у потомков гибридных растений. В том специальном случае, когда некий признак встречается лишь в двух вариантах: *A* (доминантный) и *a* (рецессивный), гипотеза независимости приводит к известному закону расщепления признаков в отношении 3:1. Этот закон относится к потомкам родителей α и β с генотипом *Aa*, которые производят (в равном количестве) гаметы с признаком *A* и с признаком *a*. Возможны четыре равновероятные варианта сочетания гамет, полученных от родителей, а именно: *AA, Aa, aA* и *aa*. В силу доминирования три первые варианта дают (в фенотипе потомка) доминантный признак, так что вероятность этого есть p= ¾, а последний вариант дает рецессивный признак (и вероятность этого есть q= ¼). Отношение вероятностей $p:q$ равно 3:1, что и называется законом расщепления признаков в отношении 3:1. Следовательно, вероятность получить среди n потомков m особей с признаком *A* задается биномиальным законом

$$P_n(m) = C_n^m p^m q^{n-m}, \quad p=3/4, \quad q=1/4 \qquad (1)$$

Этот закон и представляет собой ту статистическую гипотезу (или модель) *H*, которая подлежит проверке по экспериментальным данным.

2.2. Принципы обработки фактических данных.

Отметим что, чем более подробно представлены исходные данные, тем интереснее их обрабатывать. Например, в опытах с горохом всего интереснее были бы данные по отдельным бобам (стручкам). Получилась бы большая таблица со строчками вида: «растение номер *i*, стручок номер *j*: всего горошин *n*, из них желтых *m*». Но обработка таких данных в докомпьютерную эпоху была бы достаточно трудоемка, поэтому данные в аналогичных экспериментах обычно суммировались по отдельным растениям, либо даже по определенным группам растений (например, по «семьям» или «семействам»).

Пусть предъявлены результаты для r семейств численностью $n_1, n_2, ..., n_r$ потомков, из которых соответственно $m_1, m_2, ..., m_r$ обладают признаком *A*. Колмогоров ставит вопрос: *«Как возможно полнее проверить, согласуется ли такой результат опыта с менделевскими допущениями или нет?»* (с.212) и дает на него следующий ответ:

Если число особей в каждом семействе очень мало (например, меньше 10), то целесообразно непосредственно проверять формулу (1) при помощи χ^2 – критерия Пирсона.

Если каждое семейство сравнительно многочисленно, то целесообразнее другой метод. В этом случае из (1) следует, что нормированные уклонения

$$x_n(m) = (m - np)/\sqrt{npq} \qquad (2)$$

подчиняются приближенно закону Гаусса с единичной дисперсией. (с.212).

[*] Здесь хотелось бы подчеркнуть, что о генетике говорится лишь как о хорошем первом приближении к действительному положению вещей. Это важно потому, что точность в первом приближении и статистическая значимость отклонений от модели – это далеко не совпадающие понятия. Например, если теоретическая вероятность обнаружить доминантный фенотип равна 0,75, а частота его обнаружения в опыте составила 0,77, то при большом числе опытов, которые характерны для генетики, это различие может быть статистически значимым. Но в качестве первого приближения число 0,75 может быть вполне приемлемым.

(В формуле (2) мы изменили обозначения Колмогорова). Это место работы [1] требует комментария. Первый вариант проверки (с помощью критерия χ^2) в работе не применялся, так что нет возможности с полной определенностью сказать, какой именно из вариантов этого критерия имеется в виду. Но, скорее всего, речь идет о следующем.

Пусть представлены данные по большому количеству семейств, каждое из которых в отдельности немногочисленно (например, для гороха могли бы быть представлены данные по отдельным бобам: сколько в каждом бобе было горошин с доминантным и рецессивным фенотипом). Сначала такие данные группируются по значениям общего числа n потомков в семействе (например, берутся данные по тем бобам, которые содержат $n=n_0=10$ горошин). Для каждого семейства с данным $n=n_0$ число m потомков с доминантным признаком может принимать (n_0 +1) значение от 0 до n_0, причем вероятности этих значений определяются формулой (1). Иначе говоря, каждому семейству отвечает одно полиномиальное испытание с числом исходов (n_0 +1), причем вероятности любого исхода известны. Критерий χ^2 (как критерий асимптотический) может применяться в том случае, когда число N полиномиальных испытаний (т.е. число семейств с данным $n=n_0$) достаточно велико. (Принято выдвигать условие $N \cdot P_{n_0}(m) \geq 5$ для всех m=0,1,..., n_0.) Если это верно, то рассматриваются количества семейств $\mu_0, \mu_1, ..., \mu_{n_0}$, которые содержат соответственно 0,1,...,n_0 потомков с доминантным признаком. Затем образуются нормированные величины $[\mu_m - N \cdot P_{n_0}(m)] / \sqrt{N \cdot P_{n_0}(m)}$, сумма квадратов которых должна (приблизительно) иметь распределение χ^2 с n_0 степенями свободы.

Заключительный этап обработки обычно состоит в том, что суммируются все значения χ^2 и степени свободы по всем возможным n_0 и оценивается статистическая значимость полученного результата.

Определенный недостаток такого варианта критерия χ^2 заключается в том, что для каждого отдельного семейства не изучается отклонение полученной численности m доминантного признака от теоретического идеала n_0p. Учитываются лишь объединенные данные $\mu_0, \mu_1, ..., \mu_{n_0}$ о численностях семейств с m=0,1,...,n_0. Поэтому переход к нормированным отклонениям (2), когда для каждого семейства вычисляется свое нормированное отклонение m от n_0p, означает более полный учет исходной информации, чем в случае применения критерия χ^2. Это формулируется в тексте работы [1] следующим образом:

Зато рассмотрение большого числа семейств средней величины дает возможность гораздо более тонкой проверки менделевских допущений при помощи рассмотрения распределения уклонений. (с.212).

После этого замечания в работе [1] начинается конкретный анализ экспериментальных данных из работы Н.И.Ермолаевой [10]. Для начала разъясняется, что $P\{|x_n(m)| \geq 1\} \approx 0{,}32$, а частота этого события в таблицах Ермолаевой очень близка к данному значению вероятности. Затем приводятся два графика эмпирических функций распределения для нормированных уклонений, построенные по данным Ермолаевой, которые визуально сравниваются с нормальной кривой. Эти графики построены нами заново (см. рис.1*а* и 1*б* в следующем пункте статьи). При одном взгляде на эти рисунки становится ясным, что имеется некое разумное согласие с нормальным законом, однако этот вывод не мешало бы проверить с помощью какого-либо статистического критерия. Применяется критерий Колмогорова, и расхождение оказывается незначимым. Вывод Колмогорова:

Материал этот, вопреки мнению самой Н.И.Ермолаевой, оказывается блестящим новым подтверждением законов Менделя (с.210*).*

Статья [1] заканчивается кратким упоминанием о результатах Т.К.Енина (работы [12], [13]).

3. Новая обработка данных Н.И.Ермолаевой

Причиной, побудившей нас произвести новую, более тщательную обработку материалов Ермолаевой, явилось следующее обстоятельство. Дело в том, что в таблицах Ермолаевой многие численности семейств совсем невелики (n_i менее 10). Поэтому распределения нормированных уклонений $x_n(m)$ (см. формулу (2)) могут значительно отличаться от нормального. Фактически речь идет о промежуточном случае между двумя вариантами, указанными в [1]: первый - малочисленные семейства (n_i невелики) и критерий χ^2; второй - большие n_i и критерий Колмогорова. Интересно выяснить, в какой мере это обстоятельство влияет на выводы работы [1].

3.1. Экспериментальные данные Ермолаевой и их обработка.

Как известно, Н.И.Ермолаева была аспиранткой Т.Д.Лысенко, и перед ней была поставлена задача – доказать, что законы Менделя не подтверждаются экспериментом. В теоретическом отношении работа [10]

безграмотна: автор считает, что число успехов m в n биномиальных испытаниях всегда должно удовлетворять неравенству $|m - np| \le \sqrt{npq}$. (На самом деле, вероятность нарушения этого неравенства равна примерно 1/3, на что, в частности обращается внимание в [1].) Но эта безграмотность имела ту положительную сторону, что Ермолаева всегда могла найти много «нарушений» закона Менделя, так что обстоятельства, в которых она работала, видимо, не вынуждали ее к какой-то систематической фальсификации результатов.

В работе [10] приведены результаты двух экспериментов по скрещиванию различных сортов гороха по «менделирующим» признакам. Первый эксперимент – исследование 98 семей по признаку «окраска пазушного кольца и цветка» (белый/красный), второй – исследование 122 семей по признаку «окраска семядолей» (желтая/зеленая). Более подробное описание экспериментов см. в Приложении 1. Здесь же необходимо сказать, что, к сожалению, результаты представлены крайне небрежно, с большим количеством явных ошибок. По этой причине мы решили повторить обработку представленных в [10] экспериментальных данных, исправив очевидные ошибки.

Построенные нами эмпирические функции распределения для нормированных отклонений $x_n(m)$ приведены на рис.1*а* и 1*б*.

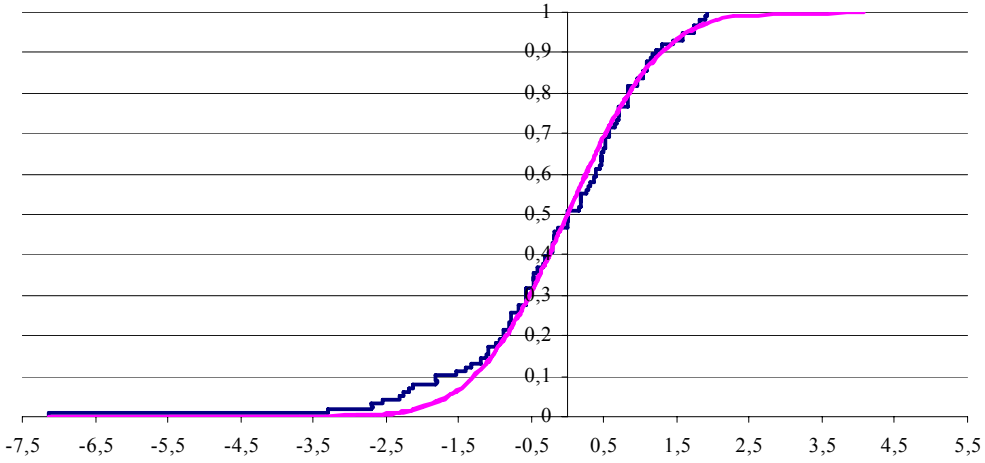

Рис. 1*а* . Данные Ермолаевой (табл.4 из [10]): обычный масштаб

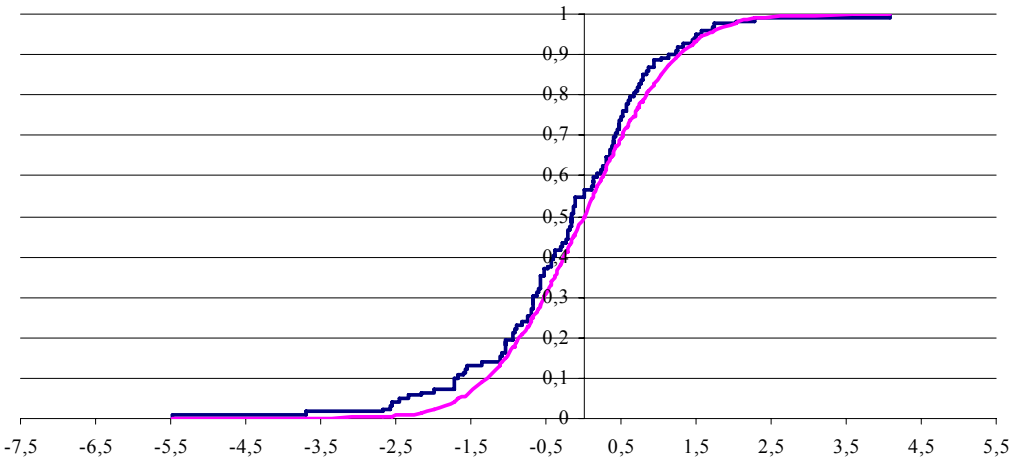

Рис. 1*б* . Данные Ермолаевой (табл.6 из [10]): обычный масштаб

Графики на этих рисунках визуально схожи с соответствующими графиками из [1], но не являются точным их повторением. Поскольку при компьютерных вычислениях резко снижается вероятность ошибок, то логично предположить, что именно в работе [1] были допущены небольшие погрешности в вычислениях.

Значения статистики Колмогорова $\lambda = \sqrt{n}\, D_n$, вычисленные по соответствующим данным, таковы: для первого эксперимента (98 семей) $\lambda = 0{,}66$ (у Колмогорова 0,82), для второго эксперимента (122 семьи) $\lambda = 1{,}00$ (у Колмогорова 0,75). Впрочем, без всяких изменений сохраняется вывод о статистической незначимости этих показателей: $K(0{,}66) = 0{,}22$, $K(1{,}00) = 0{,}73$. Соответствующие p-значения получаются вычитанием из 1, т.е. равны, соответственно, 0,78 и 0,27 (и не близки к 0 или 1).

3.2. Применение нормального масштаба.

При исследовании нормальности выборок часто применяется так называемый «нормальный» масштаб. Это означает, что вместо оси ординат *Y* (по которой откладываются вероятности, либо частоты) используется ось $Z = \Phi^{-1}(Y)$, где Φ - функция Лапласа. В этом масштабе стандартное нормальное распределение изображается биссектрисой координатного угла. Вместо ступенчатой функции распределения, однако, чаще изображают середины скачков этой функции, соединяемые отрезками прямых.

На рис.2*а* и 2*б* мы изобразили те же данные, что на рис.1*а* и 1*б*, но в нормальном масштабе.

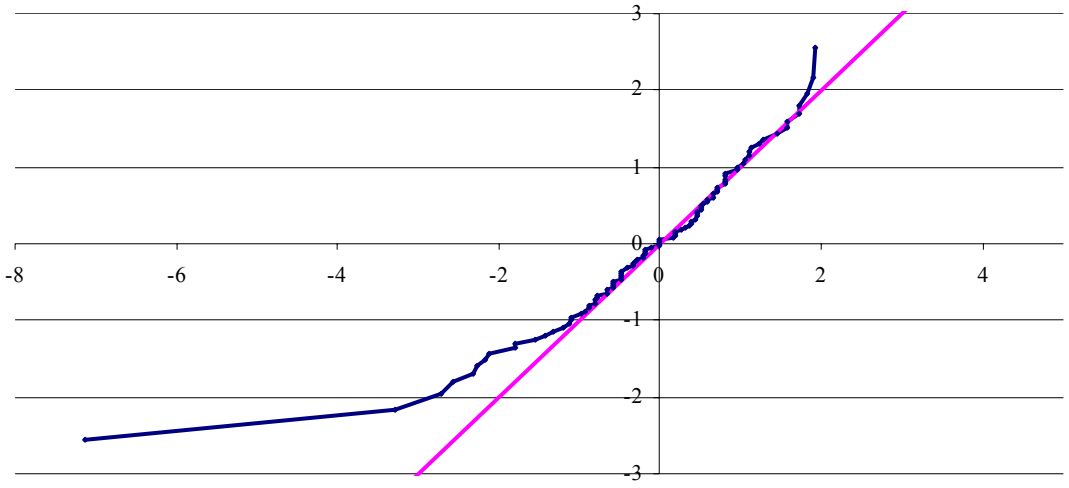

Рис.2*а* . Данные Ермолаевой (табл.4 из [10]): нормальный масштаб

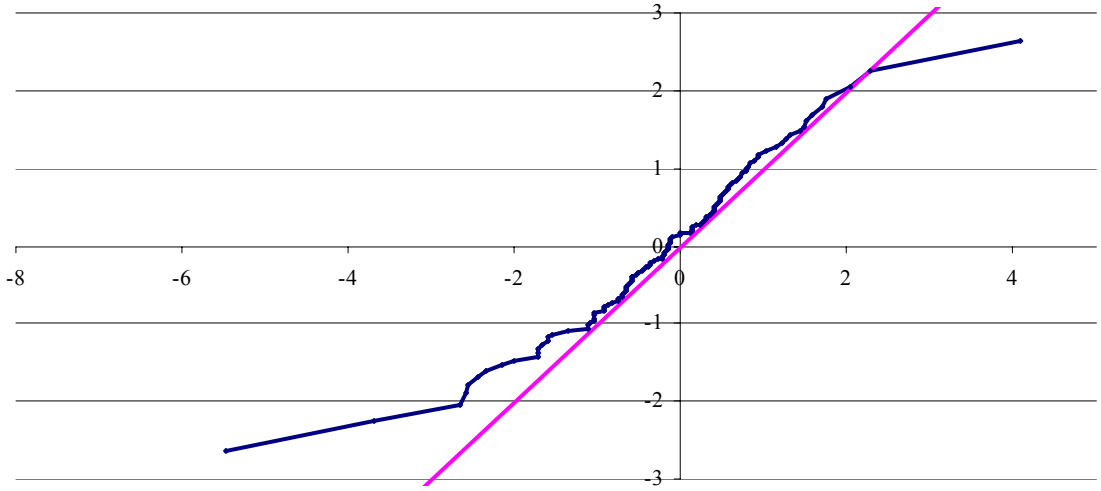

Рис.2*б* . Данные Ермолаевой (табл.6 из [10]): нормальный масштаб

Видна типичная картина «тяжелого левого хвоста»: большие отрицательные значения величин $x_n(m)$ встречаются с гораздо большей частотой, чем полагалось бы по нормальному закону. (Им соответствуют меньшие по модулю отрицательные значения *Z* и, следовательно, большие вероятности.)

Глядя на рисунки 2*а* и 2*б*, ни один статистик не согласится с тем, что на них представлены выборки из нормального закона. Обычный масштаб (рис.1*а* и 1*б*) затушевывает различия в хвостах распределений, а нормальный масштаб (рис.2*а* и 2*б*) их проявляет. При желании можно найти и статистический критерий, который отвергнет нормальность. Таков, например, критерий Реньи (см. [9], стр.82).

Он основан на статистике

$$R_n^+(a,1) = \sup_{x:F(x)>=a} \{(F_n(x)-F(x))/F(x)\}.$$

Иными словами, при применении этого критерия кроме вида самой статистики $(F_n(x)-F(x))/F(x)$ можно выбирать еще значение *a*, ограничивающее область изменения *x*. Например, если выбрать $a = 0{,}02$, то

для данных рис.1*а* получаем $R^+ = 1{,}886$, а для рис.1*б* $R^+ = 2{,}243$. Соответствующие *p*-значения, найденные по рекомендациям из [9], составляют 0,0076 и 0,0004, соответственно, так что гипотеза нормальности отвергается.

Поистине, «статистика – что дышло …»: критерий Колмогорова не отвергает нормальность, а критерий Реньи – отвергает. Чего же требует в таком случае та «моральная добропорядочность», которую следует иметь статистику?

Прежде всего, нужно исследовать, нет ли математических причин для появления таких «хвостов» распределения. Эти причины могут быть связаны, например, с малочисленностью семей, представленных экспериментальными данными из работы [10].

3.3. Математические уточнения.

При малом *n* величина $x_n(m)$ не имеет нормального распределения. А эти величины, подсчитанные для разных семей (с несовпадающими численностями), вообще не образуют выборки, поскольку их распределения различны. Однако, не запрещается рассматривать аналог эмпирической функции распределения для *r* семейств

$$F_r(x) = \{\text{число } x_i<x\}/r \qquad (3)$$

«Теоретическим идеалом» для функции $F_r(x)$ является ее математическое ожидание $\mathbf{E}F_r(x)$. Прав ли Колмогоров в том, что предлагает считать «теоретическим идеалом» функцию Лапласа $\Phi(x)$?

Для ответа на этот вопрос можно сосчитать (с помощью компьютера) $\mathbf{E}F_r(x)$ для заданного набора численностей семей $n_1,n_2,\ldots,n_r$ (того, который отвечает таблице 4 или 6 работы [10]). Для этого нужно лишь уметь вычислять (для значений *x* с шагом, скажем, 0.01) вероятности $P\{x_{n_i}(m_i)<x\}$, что сводится к распределению вероятностей для m_i, т.е. к биномиальному закону. Результаты расчетов (в нормальном масштабе) представлены на рис.3*а* и 3*б*.

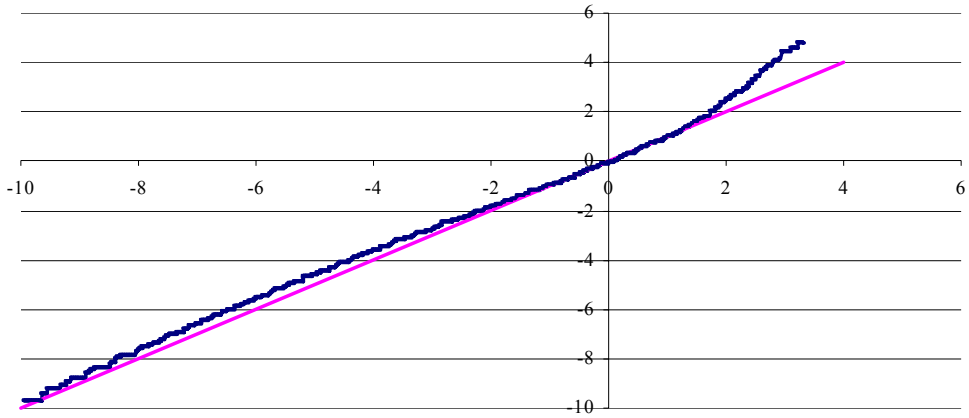

Рис.3*а*. Теоретический идеал для таблицы 4 из [10].

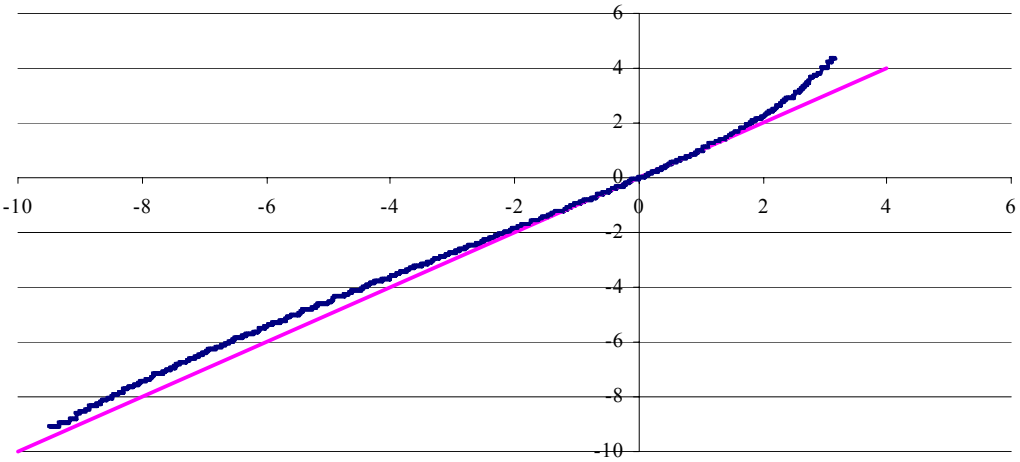

Рис.3*б* . Теоретический идеал для таблицы 6 из [10].

Из этих рисунков видно, что в качественном отношении $\mathbf{E}F_r(x)$ отклоняется от нормального закона

примерно так же, как эмпирические функции на рис.2*а* и 2*б*. Но эти отклонения заметны в области $|x| > 2$, а на рис.2*а* и 2*б* отклонение левого хвоста больше и проявляется при $x < -1$. (В области же $|x| > 2$ на рис.2*а* и 2*б* очень мало наблюдений). Таким образом, в области $|x| < 2$ подход Колмогорова совершенно правилен: несмотря на малые численности семей, теоретический идеал совпадает с нормальным законом.

Тем не менее, возникает вопрос о том, как велики могут быть отклонения полученной в эксперименте функции $F_r(x)$ от предполагаемого теоретического идеала, т.е. о распределении статистики $\lambda = \sqrt{n}\ \sup_x |F_r(x) - \Phi(x)|$. Другими словами, мы хотим выяснить, насколько правомерно было применять критерий Колмогорова, т.е. какое распределение имеет λ для нашего набора $n_1,\ldots,n_r$ и насколько сильно оно отличается от $K(y)$. Эти распределения для таблиц 4 и 6 (из работы [10]) мы сосчитали методом Монте-Карло, употребляя 10000 реализаций. Отдельная реализация состояла в том, что для данного набора n_i моделировались численности m_i, вычислялась функция $F_r(x)$ и значение статистики λ. Эмпирическая функция распределения 10000 значений λ рассматривалась, как практически точное приближение к истинному распределению. Эти эмпирические функции вместе с функцией Колмогорова $K(x)$ представлены на рис.4.

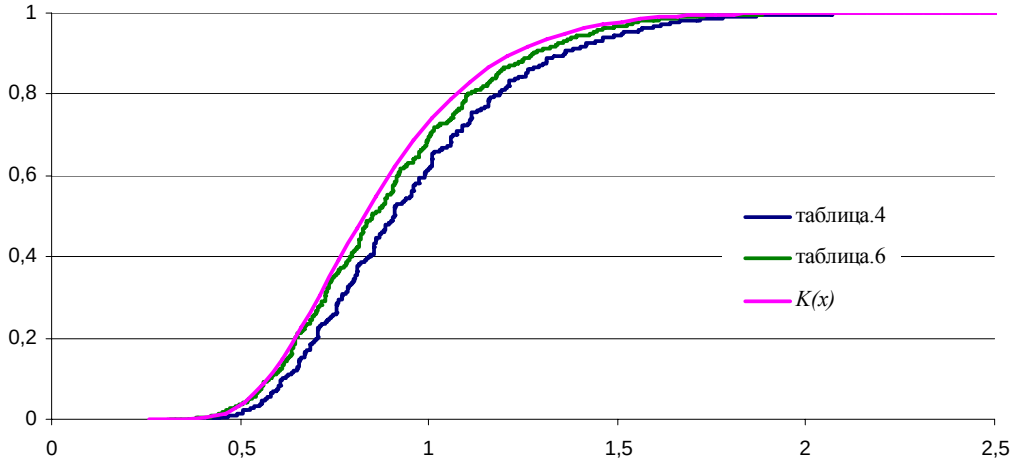

Рис.4. Функция Колмогорова и моделированные распределения статистик Колмогорова λ для таблиц 4 и 6 из [10].

Различие между функцией Колмогорова и функциями, полученными методом Монте-Карло, довольно заметно. Для таблицы 4 (98 наблюдений) отличие от функции $K(x)$ несколько больше, чем для таблицы 6 (122 наблюдения), что объясняется тем, что семьи в табл.4 более малочисленны, чем в табл.6. Однако обе эмпирические функции лежат правее функции $K(x)$, т.е. допустимые значения для статистики λ оказываются больше, чем в распределении Колмогорова. Иными словами, использование функции Колмогорова может лишь завысить статистическую значимость отклонений, но не влияет на вывод в случае незначимости.

Таким образом, сделанные математические уточнения не позволяют объяснить поведение хвостов распределений на рисунках 2*а* и 2*б*. В таких случаях необходимо обращаться к дополнительному экспериментальному материалу. К счастью, в отношении данных Ермолаевой такая возможность имеется (работа [11]). Подробное рассмотрение этих данных (см. Приложение 1) наводит на мысль о недостаточно тщательно подготовленном и проведенном эксперименте. Сводные результаты обработки данных в виде рисунков, аналогичных рис.1*а* и 1*б* (отдельно для расщепления семян красноцветковых и белоцветковых растений) приводятся на рис.5*а* и 5*б*.

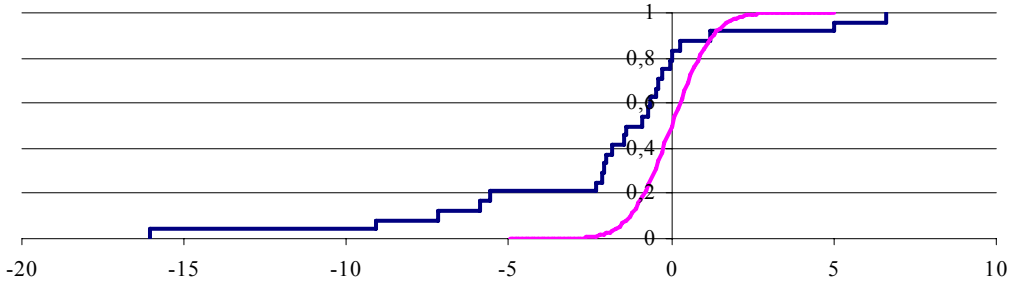

Рис. 5*а* . Данные Ермолаевой (табл.1 из [11]), красноцветковые растения из F_1.

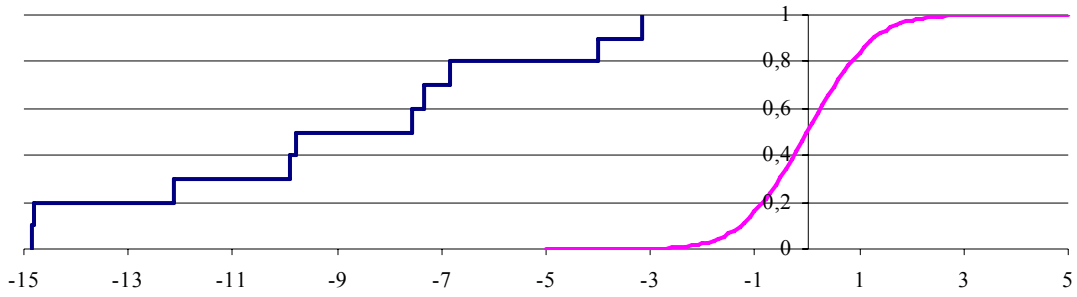

Рис. 5*б* . Данные Ермолаевой (табл.1 из [11]), белоцветковые растения из F_1.

Изображенные на этих рисунках эмпирические функции распределения не имеют ничего общего с нормальной кривой.

В этой более ранней работе Ермолаева просто не сумела воспроизвести менделевскую ситуацию в эксперименте. Но сколь благодетельна оказалась аспирантура, хотя бы и под руководством Т.Д.Лысенко! В отличие от рисунков 5*а* и 5*б*, рисунки 1*а* и 1*б* показывают значительный прогресс в умении экспериментировать. Тяжелые хвосты распределений на рис.2*а* и 2*б* в такой ситуации, вероятно, следует списать на то, что техника экспериментирования освоена еще не до конца. Говорить об установлении реальных нарушений закона Менделя на основании этих хвостов вряд ли стоит.

Математические уточнения методики статистической обработки, подробно описанные в этом параграфе (которые стали возможными благодаря использованию компьютера) лишь подтверждают правильность подхода А.Н.Колмогорова, несмотря на малочисленность многих экспериментальных семейств. Единственное, в чем с ним можно не вполне согласиться, это в том, что данные Ермолаевой являются «блестящим подтверждением» законов Менделя. Подтверждение действительно есть, но довольно скромное: скорее всего исходные материалы и/или техника эксперимента еще не доведены до нужного совершенства.

4. Обработка данных Т.К.Енина.

В списке литературы к работе Колмогорова [1] упомянуты две работы Т.К.Енина ([12] и [13]), но в самом тексте работы сказано несколько слов лишь об одной из них ([12]). Приведем соответствующую цитату.

Если бы в какой-либо достаточно обширной серии семейств уклонения m/n от 3/4 были бы систематически меньше, чем требует теория, то это в такой же мере опровергало бы применимость к этой серии семейств, сформулированных выше допущений, как и систематическое превышение теоретически предсказываемых размеров этих уклонений. Намек на такую систематическую чрезмерную близость частот m/n к 3/4 имеется в материалах работы Т.К. Енина (1). Однако материалы этой работы недостаточно обширны (25 семейств по сравнению с двумя сериями в 98 и 123** семейства у Н.И. Ермолаевой) и возбуждают ряд других сомнений (сам автор считает их не вполне однородными). Поэтому в детальное их рассмотрение мы входить не будем.* (с.214).

К этой цитате нужно сделать следующий комментарий. Нетрудно себе представить ситуации (и Колмогоров приводит их в начале работы [1]), в которых вероятность проявления доминантного признака может быть отличной от 3/4 Но крайне трудно (и до сего времени не удалось: см. ниже п.5) предложить такую модель расщепления признаков, которая могла бы повлечь бóльшую близость частот m/n к вероятности 3/4, чем та, которая вытекает из биномиального закона для m (при фиксированном n). Таким образом, при наблюдении слишком большой близости ставится под сомнение не теория, а та самая «моральная добропорядочность», о которой сказано во введении, но на этот раз не статистика, а экспериментатора. Проще говоря, выдвигается подозрение в фальсификации результатов наблюдений (например, в том, что публикуются результаты не всех наблюдений, а только тех, для которых m/n оказалось особенно близким к 3/4). В данном параграфе статьи мы рассмотрим этот упрек по отношению к данным Енина, а в следующем п.5 – по отношению к данным самого Менделя.

В работе [12] речь идет о расщеплении в поколении F_2 у томатов по признаку формы листьев (доминантный признак - нормально рассеченные листья, рецессивный признак – картофелевидные листья) (см. Приложение 2). Неоднородность данных связана с тем, что семена растений F_1 высевались для получения растений F_2 в два срока: 10 семейств получены при более раннем сроке посева, 15 семейств при более позднем. По мнению автора, растения более раннего срока посева страдали от недостатка тепла и

* Работа[12] по нашему списку.
** На самом деле 122.

света, что могло повлиять на результаты расщепления признаков. Семейства Енина гораздо более многочисленны, чем семейства Ермолаевой (до нескольких сотен особей), так что нормальность распределения нормированных отклонений $x_n(m)$ сомнений не вызывает (при условии соблюдения закона (1)).

Посмотрим, как ведут себя p-значения при проверке гипотезы H_0, состоящей в том, что теоретическая частота появления доминантного признака равна 3/4 (альтернативная гипотеза H_a: теоретическая частота не равна 3/4). В качестве статистики критерия можно взять величину $T=| x_n(m) |$, тогда p-значение $=P\{x>T_{выб}\}$. Результаты такой проверки приведены на рис.6.

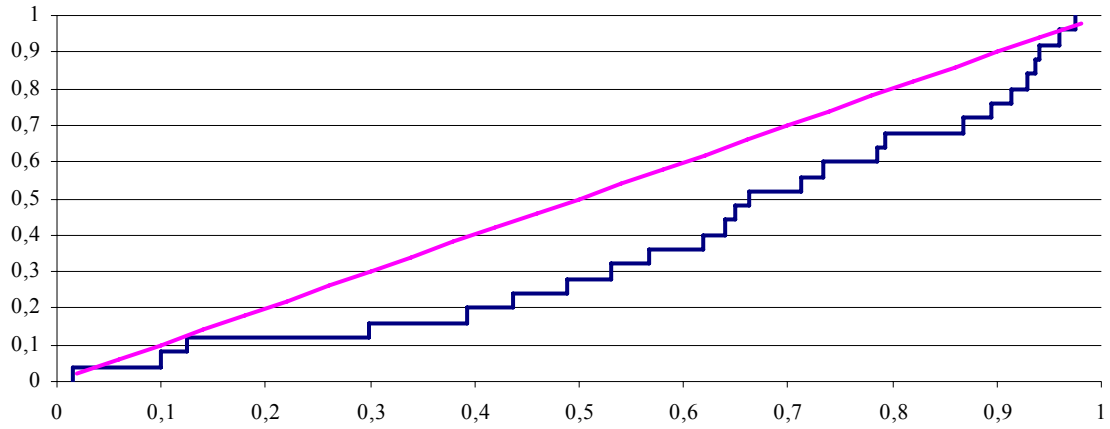

Рис. 6. Распределение p -значений для данных Енина [12].

Как было сказано выше (см.п.1), если верна нулевая гипотеза, то при неоднократной ее проверке p-значения образуют выборку из равномерного распределения. На рис.6 видно, что p-значения далеки от ожидаемых и смещаются в сторону увеличения, т.е. наблюдаемая частота ближе к 3/4, чем должна быть в предложенной биномиальной схеме. Это наводит на мысль о возможности тенденциозной выбраковки данных эксперимента.

Рассмотрим эмпирические функции распределения отклонений $x_n(m)$ представленные на рис.7 в нормальном масштабе отдельно для 10 и 15 наблюдений.

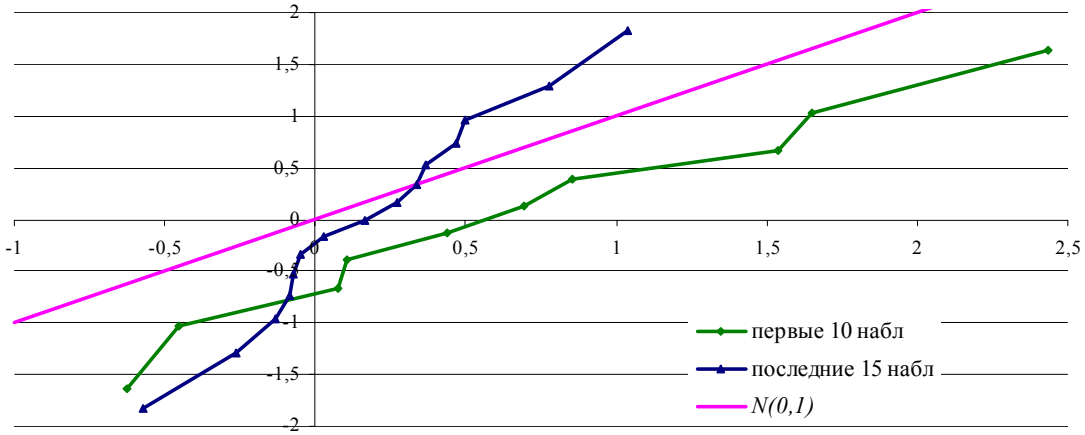

Рис.7. Данные Енина [12].

Они совсем не похожи на стандартный нормальный закон, представленный на этом рисунке биссектрисой координатного угла. Первые 10 наблюдений имеют явный систематический сдвиг, а вторые 15 – меньший сдвиг, но больший, чем биссектриса наклон. Соответствующие статистики для первой группы $\bar{x}$ =0,671, s=0,973; для второй группы $\bar{x}$ =0,185, s=0,416. Используя стандартные приемы статистики, нетрудно показать, что отклонения от стандартного нормального распределения высоко значимы. Формальное применение статистической проверки гипотез дает, таким образом, тот вывод, что в опытах Енина законы Менделя нарушены.

Оценим, однако, насколько велики эти нарушения. Суммарно в первой группе (10 наблюдений) всего n=2678 растений, из них с доминантным признаком 2059, так что отношение m/n=0,7689≈0,77 вместо теоретического значения 0,75. Отклонение невелико (хотя и статистически значимо из-за большого значения n). Здравый смысл в данном случае вступает в противоречие со статистикой и говорит, что имеет место

некоторое скромное подтверждение законов Менделя. Что касается явно низкого значения s=0,416 для второй группы наблюдений (теоретическое σ=1) то его объяснить нечем. Как и в случае с Ермолаевой, желательно обратиться к анализу других данных того же автора (См. Приложение 1).

Простейший прием суммирования всех данных в ряде случаев приводит к выводу, что результаты опытов Енина статистически значимо отклоняются от теоретических вероятностей. Но автор не хочет замечать этих различий, и в этом смысле некритичен по отношению к проверяемой теории. С другой стороны, эти суммарные различия невелики (хотя и значимы). Тестирование величины отклонений от «теоретического идеала» почти во всех случаях показывает слишком малую величину отклонений. Это обстоятельство, а также распределение p-значений при неоднократной проверке соответствующей гипотезы трудно объяснить иначе, как не вполне корректной публикацией полученных в эксперименте результатов.

5.Споры вокруг данных Менделя.

Тот же упрек, который А.Н Колмогоров предъявил Т.К.Енину – в чрезмерной близости наблюдений к теоретическому идеалу, в течение многих десятков лет предъявляется и самому Г. Менделю. Согласно [14], первое предъявление этого упрека было сделано еще в 1902 году Уэлдоном, но не привлекло большого внимания. Однако в 1936 году за это дело весьма основательно взялся Р.Фишер [6], который восстановил, насколько это возможно в деталях, ход экспериментов Г.Менделя. По оценкам Фишера Мендель должен был одновременно выращивать несколько тысяч растений гороха, которые с трудом могли поместиться на имевшемся у Менделя участке земли (не говоря уже о затратах труда на сбор и анализ результатов опытов). Мендель, стало быть, должен был экономить на объеме экспериментов. С этим связан следующий (поистине убийственный) аргумент Фишера.

В числе опытов Менделя были такие, в которых он выяснял генотипы отдельных растений из F_2 , которые фенотипически обладали доминантным признаком A. При этом возможен либо генотип AA, либо генотип Aa, причем отношение их численностей теоретически равно 1:2. Это соотношение и было предметом проверки. В том случае, когда признак $A(a)$ относится к семенам, достаточно исследовать семена растения. Но в том случае, когда признак относится к целому растению, нужно вырастить из его семян некоторое количество растений и посмотреть, не проявляют ли некоторые из них рецессивного фенотипа a. В работе Менделя [15] есть фраза, которую можно понять так, что он выращивал ровно 10 таких тестовых растений, и если ни одно из них не проявляло признака a, то считал, что исходное растение имеет генотип AA. Фишер справедливо замечает, что в случае генотипа Aa вероятность получить 10 потомков с фенотипом A составляет $(3/4)^{10}$=0,0563, т.е. не слишком малую величину. Следовательно, теоретическое соотношение при такой постановке эксперимента должно быть не 1:2, а (1+2·0,0563):2(1-0,0563)=1,1126:1,8874. Между тем Мендель сообщает, что для 600 растений получил отношение 201:399, что весьма близко к 1:2 (даже подозрительно близко), но довольно далеко от отношения 222,5:377,5, которое в данном случае представляет собой теоретический идеал.

Рассматривая все данные Менделя и вычисляя каждый раз статистику χ^2 (а затем складывая все значения статистики и все степени свободы, как это обычно и делается) Фишер приходит к выводу, что эти значения χ^2 недопустимо малы. Он предполагает, что некий помощник Менделя слишком хорошо знал от своего руководителя, что именно должно получаться в опытах, и соответственно подгонял числа. А в вышеописанной ситуации сам Мендель ошибался в том, какие числа должны получиться, и потому подгонка была произведена неправильно. (Кстати, через несколько десятков лет после написания работы [6] выяснилось, что у Менделя, в самом деле, были помощники.)

Впрочем, авторы [14] подвергают сомнению этот аргумент Фишера на том основании, что невозможно выращивать ровно 10 потомков, потому, что некоторые семена не дают всходов. Если же Мендель сеял на самом деле более 10 семян, то самый убийственный аргумент отводится. Но значения χ^2 все равно остаются недопустимо малыми.

В данной работе мы исследуем вопрос о том, что получится, если вместо критерия χ^2 применить критерий Колмогорова. Для этого часть данных Менделя нужно линейно преобразовать, что и описывается ниже.

5.1. Математические преобразования.

Данные Ермолаевой, Енина и часть данных Менделя можно назвать *биномиальными* в том смысле, что каждый объект исследования (т.е. отдельное семя, либо растение) может обладать одним из двух признаков (фенотип A или a, генотип AA или aa и т.д.) В этом случае результаты n наблюдений, m из которых обладают одним вариантом признака, приводятся к нормальному распределению преобразованием $x_n(m) = (m - np)/\sqrt{npq}$. Несколько таких нормированных отклонений, которые мы обозначим $x_1, x_2, \ldots, x_r$, согласно Колмогорову, изображаются в виде эмпирической функции распределения, которую можно сравнивать с нормальной, используя критерий Колмогорова или другие критерии (например, выше был использован критерий Реньи). Если же к отклонениям $x_1, x_2, \ldots, x_r$ применяется критерий χ^2 в том его варианте, который используется в [14] и [6], то эмпирическая функция распределения не строится, а все

наблюдения объединяются в одну статистику $\chi_r^2 = \sum_{i=1}^{r} |x_i|^2$, значение которой сравнивается с числом степеней свободы $r=\mathbf{E}\chi_г^2$. Понятно, что подход, включающий визуальный анализ эмпирического распределения, более информативен, чем подход, основанный на единственном значении $\chi_г^2$.

Но часть данных Менделя можно назвать *полиномиальными* в том смысле, что каждый объект исследования может быть отнесен к одному из $k>2$ классов (это бывает в том случае, когда объекты классифицируются по нескольким фенотипическим или генотипическим признакам). Напомним соответствующий критерий χ^2.

Пусть произведено N полиномиальных испытаний (т.е., в данном случае, рассмотрено N объектов), причем $\mu_1,\mu_2,\ldots,\mu_k$ – суммарное число наблюдений, попавших в классы $1,2,\ldots,k$ (при этом $\mu_1+\mu_2+\ldots+\mu_k=N$). Пусть проверяется гипотеза, состоящая в том, что вероятности попадания в отдельные классы равны $p_1,p_2,\ldots,p_k$ (при этом $p_1+p_2+\ldots+p_k=1$). В таком случае наблюдения $\mu_1, \mu_2, \ldots, \mu_k$ нормируются несколько иначе, чем в случае $k=2$, а именно, составляются величины

$$y_1 = \frac{\mu_1 - Np_1}{\sqrt{Np_1}}, \quad y_2 = \frac{\mu_2 - Np_2}{\sqrt{Np_2}}, \ldots, y_k = \frac{\mu_k - Np_k}{\sqrt{Np_k}} \tag{4}$$

(а не величины $x_j = (\mu_j - Np_j)/\sqrt{Np_j(1-p_j)}$, как в случае биномиальных данных). Далее составляется величина $\chi_{k-1}^2 = \sum_{j=1}^{k} y_j^2$, относительно которой известно, что она (асимптотически при $N\to\infty$) имеет распределение хи-квадрат с $(k-1)$ степенями свободы (при верной проверяемой гипотезе). Таким образом числа $\mu_1,\mu_2,\ldots,\mu_k$ (либо числа $y_1, y_2, \ldots, y_k$) опять-таки отдельно не анализируются, а все сводится к одному числу χ_{k-1}^2. Однако, возможно сделать линейное преобразование величин (4), которое дает $(k-1)$ независимых нормальных величин, и их можно исследовать по методике Колмогорова [1].

Это делается следующим образом. Введем обозначения:

$$p = (p_1, p_2, \ldots, p_k), \quad \sqrt{p} = (\sqrt{p_1}, \sqrt{p_2}, \ldots, \sqrt{p_k}), \quad y = (y_1, y_2, \ldots, y_k).$$

Очевидно, выполняется равенство: $\sum_{j=1}^{k} y_j \cdot \sqrt{p_j} = (y, \sqrt{p}) = 0,$ т.е. вектор y лежит в $(k-1)$-мерной гиперплоскости L, ортогональной вектору $\sqrt{p}$. Несложные вычисления показывают, что если рассмотреть матрицу ковариаций вектора y в гиперплоскости L ,то она равна единичной матрице. Кроме того, в силу многомерной центральной предельной теоремы, распределение y асимптотически нормально. Таким образом, если выбрать в гиперплоскости L какой-то ортонормированный базис $e_1,e_2,\ldots,e_{k-1}$, то величины $z_j=(e_j,y_j)$, $j=1,\ldots,k-1$, должны образовывать выборку из стандартного нормального распределения. (В частности, поскольку $(y,y) = \sum_{j=1}^{k} y_j^2 = \sum_{j=1}^{k} z_j^2$, то отсюда получается критерий χ^2.) Иными словами, вектор $\sqrt{p}$, определяемый вероятностями $p_1,p_2,\ldots,p_k$, надо дополнить векторами $e_1,e_2,\ldots,e_{k-1}$ до базиса в R^k и затем вычислить величины $z_j=(e_j,y)$. (Нетрудно проверить, что в случае $k=2$ как раз и получается $z_1 = x_1 = (m-np)/\sqrt{npq}$.) Набор векторов $(e_1,e_2,\ldots,e_{k-1})$ определяется неоднозначно, только лишь для корректности процедуры нужно сначала выбрать этот набор, и лишь затем обратиться к данным эксперимента. В принципе, можно с помощью компьютера ортогонализировать любой набор из k векторов, одним из которых является вектор $\sqrt{p}$, но мы предпочли, не используя компьютер, так выбрать векторы $e_1, e_2, \ldots, e_{k-1}$, чтобы они содержали, насколько возможно, больше нулевых координат. В этом случае величины z_j проще интерпретировать. Например, если $p_1=p_2$, то в качестве вектора e_1 можно взять $e_1=(1,-1,0,\ldots,0)/\sqrt{2}$, и величина $z_1=(e_1,y)$ окажется пропорциональной $\mu_1-\mu_2$, т.е.разности между наблюдаемыми численностями двух равновероятных исходов. Выбранные нами в конкретных случаях базисы указаны ниже. В отличие от «нормированных» величин $x_n(m)$, мы называем в дальнейшем величины z_j *«нормализованными».*

5.2. Данные Менделя и их обработка.

Результаты экспериментов Менделя мы взяли из работы [15], используя также сводную таблицу 1 работы [14]*. Все необходимые данные и результаты их обработки мы свели в четыре таблицы. Наша таблица 1 содержит только биномиальные эксперименты Менделя.

Таблица 1.

Эксперимент	Наблюдаемые численности		Соответствующие вероятности		Результаты "нормализации"
1	5474	1850	3/4	1/4	-0,5127
2	6022	2001	3/4	1/4	0,1225
3	705	224	3/4	1/4	0,6251
4	882	299	3/4	1/4	-0,2520
5	428	152	3/4	1/4	-0,6712
6	651	207	3/4	1/4	0,5913
7	787	277	3/4	1/4	-0,7788
8	372	193	2/3	1/3	-0,4165
9	353	166	2/3	1/3	0,6518
10	64	36	2/3	1/3	-0,5657
11	71	29	2/3	1/3	0,9192
12	60	40	2/3	1/3	-1,4142
13	67	33	2/3	1/3	0,0707
14	72	28	2/3	1/3	1,1314
15	65	35	2/3	1/3	-0,3536

Каждому такому эксперименту соответствует одно нормированное (оно же нормализованное) число наблюдений более вероятного признака, приведенное в последней колонке таблицы 1.

Таблица 2.

Эксперимент	Наблюдаемые численности				Соответствующие вероятности				Результаты "нормализации"		
16a	315	101			9/16	3/16			0,339		
	108	32			3/16	1/16			0,588		
									-0,098		
16b	38	60	138		1/16	1/8	1/4		1,230	0,087	
	28	65			1/16	1/8			0,615	-0,615	
	35	68			1/16	1/8			0,087	0,071	
	30	67			1/16	1/8			-0,435	-0,577	
17	8	22	45	78	1/64	1/32	1/16	1/8	-1,343	-0,671	
	14	17	36		1/64	1/32	1/16		-0,448	1,163	
	9	25	38		1/64	1/32	1/16		-0,448	-0,548	
	11	20	40		1/64	1/32	1/16		0,671	1,132	
	8	15	49		1/64	1/32	1/16		0,316	0,347	
	10	18	48		1/64	1/32	1/16		0,548	1,007	
	10	19			1/64	1/32			0,581	-0,224	
	7	24			1/64	1/32			0,791	0,112	
		14				1/32			0,791	0,237	
		18				1/32			-0,475	-1,599	
		20				1/32			-0,791	0,097	
		16				1/32			-0,633	-1,331	
									0,633	0,224	
18	20	23	25	22	1/4	1/4	1/4	1/4	-0,447	0,447	-0,422

* Заметим, что в таблице 1 имеется исправленная нами на основании [15] и [6] опечатка: для эксперимента №17 в конце второго столбца наблюденных численностей должно стоять 16 (а не 14).

19	25	19	22	21	1/4	1/4	1/4	1/4	0,910	0,152	0,107
20	31	26	27	26	1/4	1/4	1/4	1/4	0,674	0,135	0,381
21	24	22	25	27	1/4	1/4	1/4	1/4	0,286	-0,286	-0,606
22	47	40	38	41	1/4	1/4	1/4	1/4	0,768	-0,329	0,621

В таблицу 2 включены полиномиальные эксперименты. Наблюденные численности и их вероятности размещены здесь (для каждого отдельного эксперимента) в конгруентных (совмещающихся при наложении) участках таблицы. Результаты нормализации приведены в порядке номеров векторов базиса $e_1,e_2,\ldots,e_{k-1}$, которые в свою очередь описаны в таблице 3. В том случае, когда эти результаты показаны в двух столбцах, первому столбцу соответствует первая половина номеров базисных векторов, а второму столбцу – остальные.

Таблица 3.

Эксперимент	Базис		
16*a*	$e_1=\frac{1}{2}(1,\sqrt{3},0,0)$	$e_2=\frac{1}{2}(0,0,1,-\sqrt{3})$	$e_3=\frac{1}{4}(\sqrt{3},1,-3,\sqrt{3})$
16*b*	$e_1=\frac{1}{\sqrt{2}}(1,-1,0^{(7)})$,	$e_2=\frac{1}{\sqrt{2}}(0,0,1,-1,0^{(5)})$,	$e_3=\frac{1}{2}(1,1,-1,-1,0^{(5)})$,
	$e_4=\frac{1}{\sqrt{2}}(0^{(4)},1,-1,0^{(3)})$,	$e_5=\frac{1}{\sqrt{2}}(0^{(6)},1,-1,0)$,	$e_6=\frac{1}{2}(0^{(4)},1^{(2)},(-1)^{(2)},0)$,
	$e_7=\frac{1}{\sqrt{12}}(\sqrt{2}^{(4)},(-1)^{(4)},0)$,	$e_8=\frac{1}{\sqrt{48}}(1^{(4)},\sqrt{2}^{(4)},-6)$	
17	$e_1=\frac{1}{\sqrt{2}}(1,-1,0^{(25)})$,	$e_2=\frac{1}{\sqrt{2}}(0,1,-1,0^{(23)})$,	$e_3=\frac{1}{\sqrt{2}}(0^{(4)},1,-1,0^{(21)})$,
	$e_4=\frac{1}{\sqrt{2}}(0^{(6)},1,-1,0^{(19)})$,	$e_5=\frac{1}{2}(1^{(2)},(-1)^{(2)},0^{(23)})$,	$e_6=\frac{1}{\sqrt{12}}(1^{(4)},(-2)^{(2)},0^{(21)})$,
	$e_7=\frac{1}{\sqrt{24}}((-1)^{(6)},(-3)^{(2)},0^{(19)})$,	$e_8=\frac{1}{\sqrt{2}}((0)^{(8)},1,-1,0^{(17)})$,	$e_9=\frac{1}{\sqrt{2}}(0^{(10)},1,-1,0^{(15)})$,
	$e_{10}=\frac{1}{\sqrt{2}}(0^{(12)},1,-1,0^{(13)})$,	$e_{11}=\frac{1}{\sqrt{2}}(0^{(14)},1,-1,0^{(11)})$,	$e_{12}=\frac{1}{\sqrt{2}}(0^{(16)},1,-1,0^{(9)})$,
	$e_{13}=\frac{1}{\sqrt{2}}(0^{(18)},1,-1,0^{(7)})$,	$e_{14}=\frac{1}{2}(0^{(8)},1^{(2)},(-1)^{(2)},0^{(15)})$,	$e_{15}=\frac{1}{\sqrt{12}}(0^{(8)},1^{(10)},(-2)^{(2)},0^{(13)})$,
	$e_{16}=\frac{1}{\sqrt{24}}(0^{(8)},1^{(6)},(-3)^{(2)},0^{(11)})$,	$e_{17}=\frac{1}{\sqrt{40}}(0^{(8)},1^{(8)},(-4)^{(2)},0^{(9)})$,	$e_{18}=\frac{1}{\sqrt{60}}(0^{(8)},1^{(10)},(-5)^{(2)},0^{(7)})$,
	$e_{19}=\frac{1}{\sqrt{2}}(0^{(20)},1,-1,0^{(5)})$,	$e_{20}=\frac{1}{\sqrt{2}}(0^{(22)},1,-1,0^{(3)})$,	$e_{21}=\frac{1}{\sqrt{2}}(0^{(24)},1,-1,0)$,
	$e_{22}=\frac{1}{2}(0^{(20)},1^{(2)},(-1)^{2},0^{(3)})$,	$e_{23}=\frac{1}{\sqrt{12}}(0^{(20)},1^{(4)},(-2)^{(2)},0)$,	$e_{24}=\frac{1}{\sqrt{96}}(3^{(8)},(-\sqrt{2})^{(12)},0^{(7)})$,
	$e_{25}=\frac{1}{\sqrt{672}}(3^{(8)},(3\sqrt{2})^{(12)},(-8)^{(6)},0)$,	$e_{26}=\frac{1}{\sqrt{448}}(1^{(8)},(\sqrt{2})^{(12)},2^{(6)},-14\sqrt{2})$	
С 18 по 22	$e_1=\frac{1}{\sqrt{2}}(1,-1,0^{(2)})$,	$e_2=\frac{1}{\sqrt{2}}(0^{(2)},1,-1)$,	$e_3=\frac{1}{2}(1^{(2)},(-1)^{(2)})$

Примечание. Запись $a^{(b)}$ означает, что координата, равная a, повторяется подряд на b соседних позициях вектора.

Сначала кратко опишем результаты применения критерия χ^2 к данным таблиц 1 и 2. Для каждого эксперимента вычисляется (как описано в п.5.1) значение χ^2. Все вычисленные значения сведены в таблицу 4[**]. В последнем столбце даны *p*-значения, т.е. вероятности получить такое же или большее значение χ^2 (т.е. такое же или худшее согласие с тем, что получено в эксперименте). Каждое в отдельности значение χ^2 не является недопустимо близким ни к 0, ни к 1, но в целом заметна тенденция к смещению этих значений в сторону единицы.

Таблица 4.

№ эксперимента	χ^2	Степени свободы	*p*-значение
1	0,26	1	0,608
2	0,01	1	0,903
3	0,39	1	0,532
4	0,06	1	0,801
5	0,45	1	0,502
6	0,35	1	0,554
7	0,61	1	0,436
8	0,17	1	0,677
9	0,42	1	0,515
10	0,32	1	0,572
11	0,85	1	0,358
12	2,00	1	0,157
13	0,01	1	0,944
14	1,28	1	0,258
15	0,12	1	0,724
16a	0,47	3	0,925
16b	2,81	8	0,946
17	15,32	26	0,951
18	0,58	3	0,902
19	0,86	3	0,835
20	0,62	3	0,892
21	0,53	3	0,912
22	1,08	3	0,781
Σ	29,59	67	1,000

Это вызывает подозрение в том, что согласие наблюдений с теорией слишком хорошее. Согласно [14], не было найдено каких-либо биологических объяснений этого факта, хотя в настоящее время о процессе мейоза (образования гамет) известно несравненно больше того, что о нем мог предполагать Мендель. Остается, следовательно, лишь вывод о том, что обращение Менделя с данными экспериментов было не вполне корректным.

Как оценить количественно надежность такого вывода? Стандартный способ обращения с величинами χ^2, полученными для независимых экспериментальных данных состоит в том, чтобы сложить все значения χ^2 и все числа степеней свободы. Проделав эту операцию, получили последнюю строчку таблицы 4, для которой *p*-значение неправдоподобно близко к 1. Разность 1-*p* трактуется, как вероятность отклонения верной гипотезы, т.е. в данном случае, как вероятность ошибочного обвинения Менделя. При таком подсчете эта вероятность есть $2 \cdot 10^{-5}$.

Но можно поступить иначе: проверять равномерность распределения *p*-значений на отрезке [0,1], взяв данные из таблицы 4 (без последней строчки, в которой стоит сумма). Соответствующая эмпирическая функция показана на рис.8.

[**] Аналогичная таблица в работе [14] также страдает ошибками: первые 7 значений χ^2 указаны неверно (даны ровно вдвое меньшие значения).

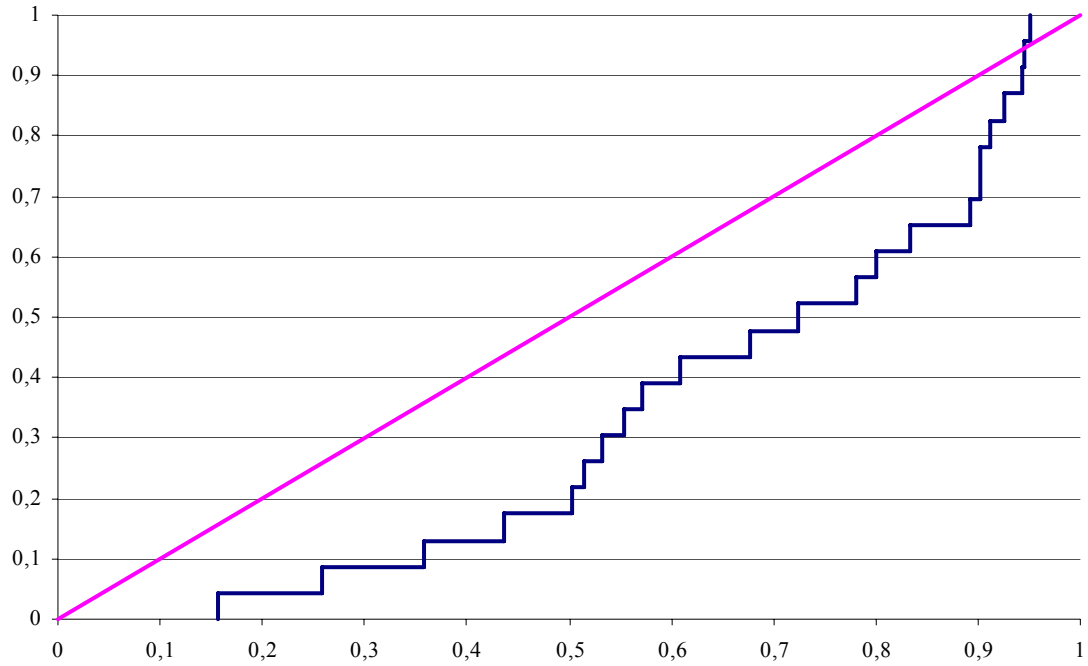

Рис.8. Распределение p -значений для данных Менделя (применение критерия χ^2)

Значение статистики Колмогорова λ = 1,573, что дает оценку для вероятности ложного обвинения 0,014. Таким образом, если исходить из значений статистики χ^2, то также можно обвинить Менделя в некорректном представлении данных эксперимента.

Теперь посмотрим, что произойдет в случае применения методики Колмогорова. Чтобы увидеть, как ведет себя эмпирическое распределение нормализованных данных по мере увеличения объема выборки, сначала берем нормализованные данные из таблицы 2 только для экспериментов 16*b* и 17. Результат графически представлен на рис.9*а*.

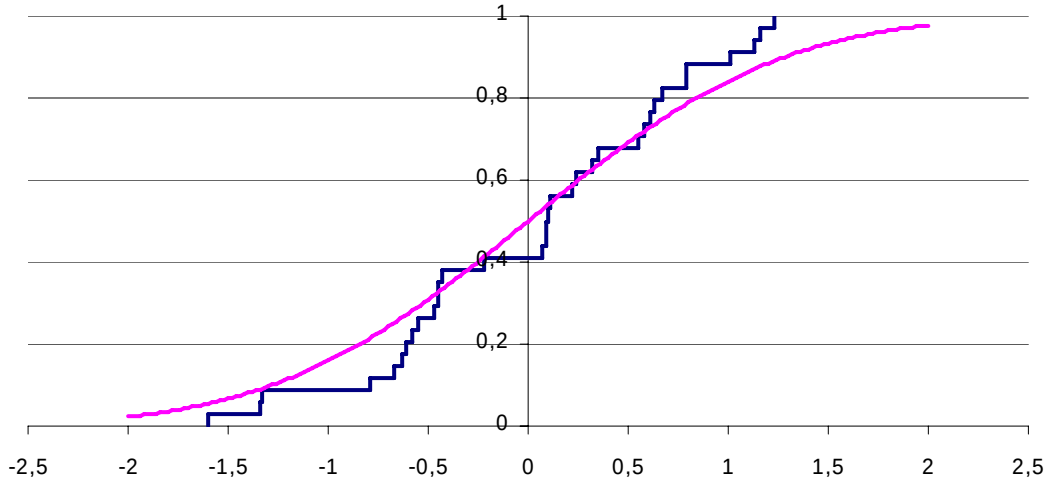

Рис. 9*а*. Нормализованные полиномиальные данные Менделя (для экспериментов 16b и 17)

Визуальное согласие с нормальной кривой неплохое, хотя заметна тенденция к большей узости эмпирического распределения в сравнении с нормальным. Значение D=0,133, что при n=34 не достигает даже 20% значимости (согласно таблице из [9]). Если же вернуться к значениям χ^2, сложив их для экспериментов 16*b* и 17, то получится 18,133 при 34 степенях свободы, что дает p=0,988, т.е. вероятность ложного обвинения ≈0,01. Критерий Колмогорова несравненно мягче относится к данным Менделя, чем χ^2.

Однако при добавлении остальных нормализованных данных из таблицы 2 визуальное впечатление слишком узкого эмпирического распределения усиливается. Все 67 нормализованных наблюдений показаны на рис.9*б*.

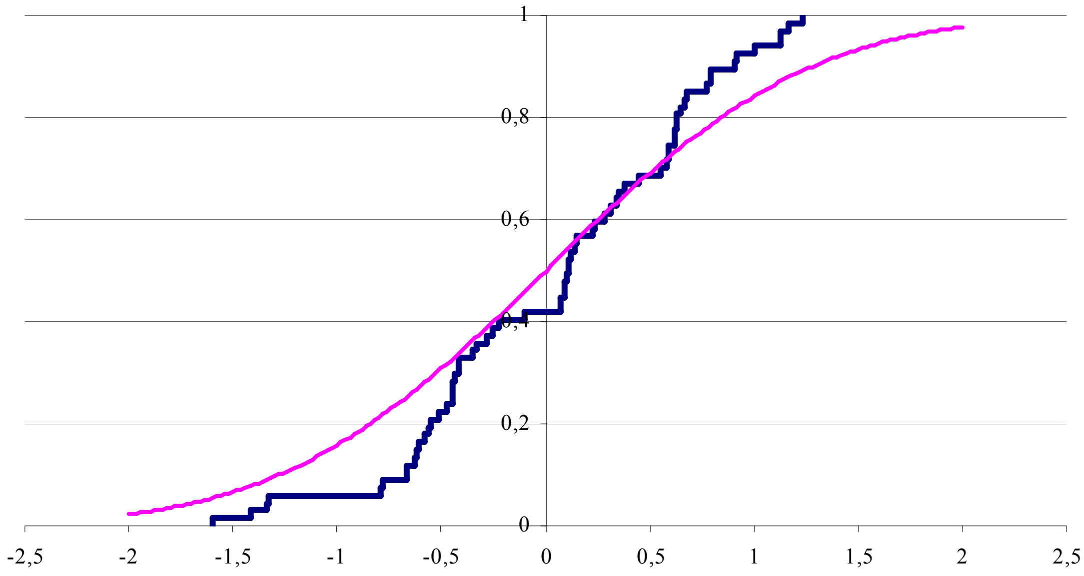

Рис.9*б* . Все нормализованные данные Менделя.

Здесь D=0,161, что значимо почти на уровне 5%.

Данные рис.9*б* приведены в нормальном масштабе на рис.10*а*.

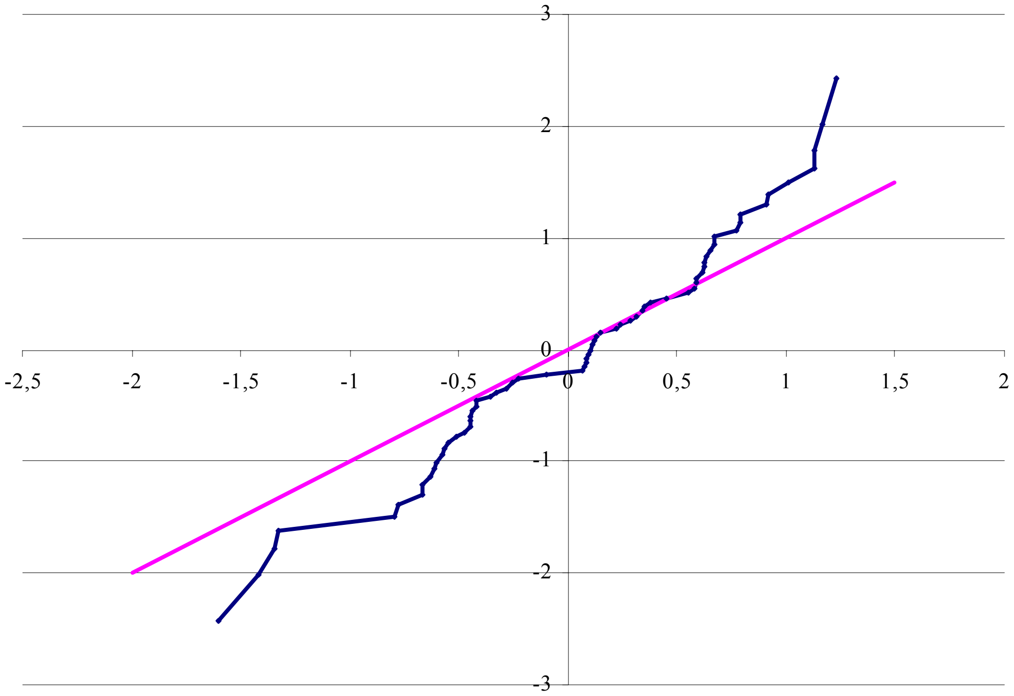

Рис. 10*а* . Все нормализованные данные Менделя в нормальном масштабе

Этот рисунок интересно сравнить с рис.10*б*, на котором (также в нормальном масштабе) изображены данные Енина (таблица 1 из работы [13]).

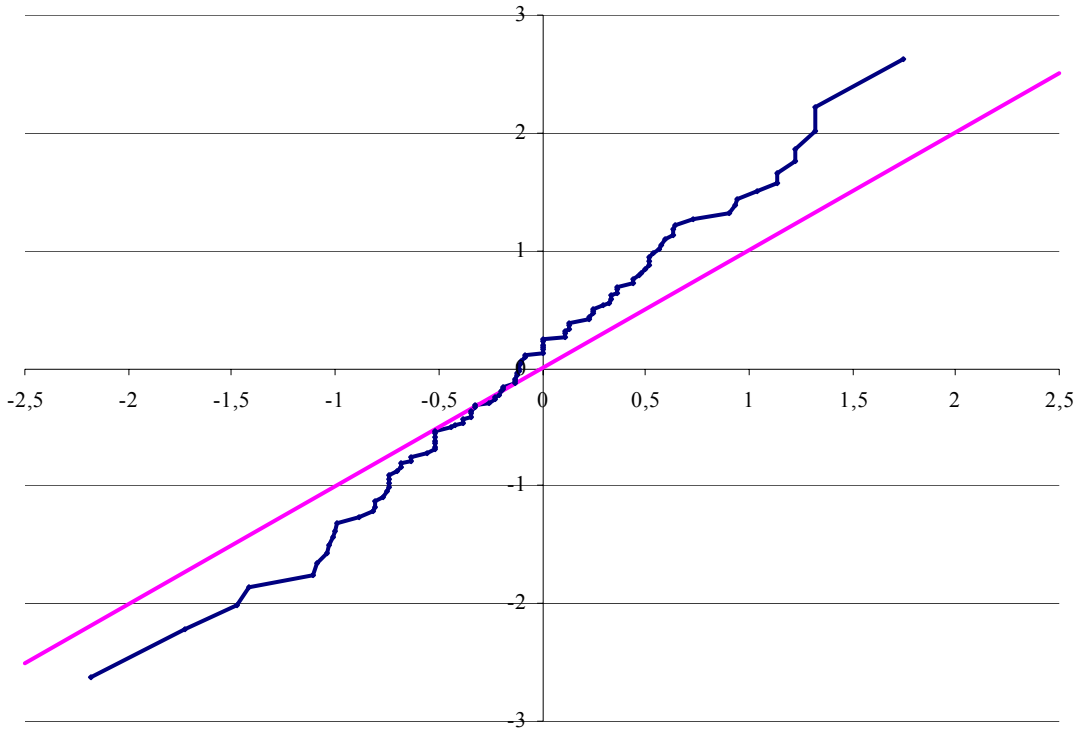

Рис. 10*б* . Данные Енина (табл 1 из [13]) в нормальном масштабе.

Рисунки очень похожи, что подтверждается и оценками стандартного отклонения: s=0,668 у Менделя и s=0,686 у Енина (вместо 1 для нормального закона).

Критерий Колмогорова дает несравненно более осторожные оценки для вероятности ложного обвинения в некорректном обращении с экспериментальными данными (чем критерий χ^2: см. [14] и [6]). Однако нельзя отрицать, что по мере объединения все большего количества нормализованных данных в одну выборку возрастает отличие выборочного распределения от нормального (в сторону большей узости эмпирического распределения). Поскольку, согласно [14], биологического объяснения этому явлению не найдено, остается принять гипотезу о какой-то форме цензуры экспериментальных данных.

6. Заключение

В данной работе мы попытались дать представление об определенной субъективности применения методов математической статистики. Рассмотренные здесь экспериментальные данные представляются нам подходящими для этой цели. Так, изображение данных Ермолаевой [10] в обычном масштабе и применение критерия Колмогорова приводит к выводу о полном согласии этих данных с законами Менделя. Если же изобразить эти данные в нормальном масштабе и применить критерий Реньи, то получается вывод о статистически значимом расхождении. Привлечение дополнительного экспериментального материала (более ранняя работа Ермолаевой [11]) позволяет разобраться в этом противоречии, предположив, что Ермолаева постепенно улучшала методику экспериментирования, но какие-то недочеты еще оставались.

Важно подчеркнуть также, что понятия статистической значимости отклонений от модели и неприменимости к реальным данным соответствующего биологического закона не тождественны. Для соответствующего примера подходят эксперименты Енина. Например, в одной из серий этих экспериментов наблюдаемая частота определенного события оказалась равной 0,77 при теоретической вероятности 0,75. Эта разница высоко статистически значима, но содержательно речь должна идти скорее о подтверждении соответствующего закона Менделя, чем об его нарушении.

Наконец, важно показать также, что статистические методы способны обнаружить не вполне корректное представление экспериментальных данных.

Этот момент иллюстрируется здесь на примере данных Т.К.Енина и самого Г.Менделя. В последнем случае показан также пример относительного произвола, связанного с выбором статистического критерия. В случае применения к данным Менделя критерия χ^2 итоговые p-значения составляют примерно $2\cdot10^{-5}$, что расценивается как убедительное доказательство не вполне корректного обращения с фактическими данными. Для того, чтобы сделать возможным применение к тем же данным другого критерия (а именно – критерия Колмогорова) нами был предложен способ линейного преобразования этих данных («нормализация»). Итоговое p-значение составляет 0,05, что гораздо менее убедительно. Однако

окончательный вывод остается тем же, что и при использовании критерия χ^2 , потому что во всех экспериментах Менделя прослеживается тенденция к лучшему согласию теории и эксперимента, чем это допускается моделью случайного сочетания гамет. Объективность методов математической статистики восстанавливается, если применять их не однократно, а к целому ряду однотипных экспериментов.

Все это важно учитывать при обучении применению методов математической статистики, иначе у исследователей часто возникает склонность приписывать получаемым выводам математическую достоверность.

Литература

1. Колмогоров А.Н. Об одном новом подтверждении законов Менделя //ДАН СССР.1940. Т 27. С. 38-42 См. перепечатку в [2].
2. Колмогоров А.Н. Теория вероятностей и математическая статистика (сб. статей) М., «Наука», 1986.
3. Тихомиров В.М. Вопросы естествознания в творчестве А.Н.Колмогорова // Вопросы истории естествознания и техники. 2003. №3.(http://www.kolmogorov.pms.ru/tihomirov-voprosy_estestvoznaniya.html)
4. Тутубалин В.Н. Теория вероятностей и случайных процессов// М. Из-во МГУ, 1992. (Стр-318-324)
5. Тюрин Ю.Н., Макаров А.А. Статистический анализ данных на компьютере. М., ИНФРА-М,1998 (стр.322-325)
6. Fisher R.A. Has Mendel's work been rediscovered? Annals of Science. 1936. V 1. (http://www.mcn.org/c/irapilgrim/men01.html)
7. Kolmogorov A.N. Sulla determinazione empirica di una legge di distribuzione. G.Ist.ital.attuar. 1933. V.4.№ 1, 83-91. Русский перевод см.[2] стр.134-141.
8. Колмогоров. Юбилейное издание в трех книгах. Книга вторая. Этих строк бегущих тесьма… Избранные места из переписки А.Н.Колмогорова и П.С.Александрова. М. Физматлит 2003. С. 436.
9. Л.Н Большев, Н.В.Смирнов. Таблицы математической статистики. М., «Наука», Физматлит, 1983.
10. Ермолаева Н.И. Еще раз о «гороховых законах».//Яровизация, 1939. №6 http://www.biometrica.tomsk.ru/misc/a12.htm
11. Ермолаева Н.И. Расщепление гороха при посеве его и скрещивании в разные сроки// Яровизация.,1938, №1-2,с. 127-134..
12. Енин Т.К. Результаты анализа расщепления гибридов томата по отдельным семьям // ДАН СССР, 1939,т. 24, с. 176-178
13. Енин Т.К. Менделизм в селекции гороха. Докл. ВАСХНИЛ,1939, №5-6, с.11-16.
14. D.J.Fairbanks, B.Rytting. Mendelian controversies: a botanical and historical review///American Journal of Botany.2001. V.88. №5. H. 737-752. (http://www.amjbot.org/cgi/content/full/88/5/737)
15. Г.Мендель Опыты над растительными гибридами. // Г.Мендель Сер. Классики науки.М.: Наука, 1965. (На англ.яз.: http://www.mendelweb.org/Mendel.html)

Приложение1. Данные Н.И.Ермолаевой.

В нашей работе мы рассматриваем 2 публикации Ермолаевой [10] и [11], упомянутые в [1].

В работе [10] описываются результаты скрещивания различных сортов гороха по различным «менделирующим» признакам: 1) окраска цветка и пазухи листа или отсутствие такой окраски (белый и красный цветок), 2) желтая и зеленая окраска семядолей. Всего в работе имеется 6 таблиц, из которых таблица 1 является сводной по всей работе, таблицы 2 и 3 являются сводными для таблиц, соответственно, 4 и 6, а таблицы 4, 5 и 6 содержат первичные результаты экспериментов. Понятно, что сводная таблица получается суммированием каких-то групп внутри таблицы первичных данных, и простая арифметика требует, чтобы результирующие суммы сходились с первичными данными. Но во многих случаях этого не происходит, частью из-за арифметических ошибок, а частью из-за того, что в сводных таблицах учтены данные, не приведенные в исходных таблицах. Таким образом, только таблицы 4 и 6 реально пригодны для обработки (таблица 5 содержит слишком мало данных).

Два первых столбца таблиц Ермолаевой носят названия «номер комбинации» и «номер семьи». По-видимому, под «номером комбинации» следует понимать данный в процессе эксперимента номер родительской пары скрещиваемых растений, в результате чего получились семена поколения F_1. При последующем посеве этих семян получаются несколько растений поколения F_1. Под семьей понимается потомство одного растения поколения F_1. Таким образом, одной комбинации скрещивания отвечает несколько семей, причем все семьи получают сплошную нумерацию. В таблице 4 сведены результаты опытов по расщеплению признака «окраска пазушного кольца и цветка», для наблюдения которого нужно вырастить растения из семян поколения F_2. Семьи в этой таблице нумеруются от 1 до 100, но две семьи (№50 и №87) пропущены, так что всего семей 98. Сопоставление чисел в каждой строке таблицы позволило нам выявить и исправить две несомненные опечатки: для семьи №22 соотношение между доминантным и рецессивным признаком исправлено с «14:4» на «12:4», а для семьи №56 аналогичный показатель исправлен с «14:2» на «4:2». Нами исправлены также арифметические ошибки в столбце «ошибка», т.е. $\sqrt{npq}$.

В таблице 6 представлено расщепление по признаку «окраска семядолей», для наблюдения которого достаточно получить семена растения F_1. Хотя комбинации скрещивания имеют те же номера, что и в таблице 4, из таблиц 2 и 3 работы [10] вытекает, что речь дет о скрещивании других родительских пар (не тех, что отвечают таблице 4). Нумерация семей идет от № 22 до № 148, но 5 семей пропущены, так что всего в таблице 6 имеется 122 семьи (не 123, как указано в [1]). Опечаток в экспериментальных данных в таблице 6 мы не обнаружили.

Таблица 4 из [10]

№ комб.	№ семьи	Ожидаемое соотн. (3: 1)		Полученное соотн.		Ошибка ±	Разница м-ду ожид. и получ.	превышение
1	1	15	5	17	3	1,94	2	превышение
	2	15	5	16	4	1,94	1	
	3	15	5	15	5	1,94	0	
2	4	16.5	5,5	11	11	2,03	5,5	превышение
	5	6.75	2,25	4	5	1,30	2,75	превышение
	6	8.25	2,75	8	3	1,44	0,25	
	7	9.75	3,25	10	3	1,56	0,25	
	8	6.75	2,25	7	2	1,30	0,25	
3	9	4.5	1,5	4	2	1,06	0,5	
	10	7.5	2,5	9	1	1,37	1,5	превышение
	11	7.5	2,5	3	7	1,37	4,5	превышение
	12	6.75	2,25	6	3	1,30	0,75	
	13	9	3	10	2	1,50	1	
	14	3.75	1,25	2	3	0,97	1,75	превышение
	15	8.25	2,75	10	1	1,44	1,75	превышение
	16	3.75	1,25	2	3	0,97	1,75	превышение
	17	4.5	1,5	4	2	1,06	0,5	
4	18	12.75	4,25	11	6	1,79	1,75	
	19	8.25	2,75	7	4	1,44	1,25	
	20	24.75	8,25	26	7	2,49	1,25	
	21	14.25	4,75	12	7	1,89	2,25	превышение
	22	12	4	12	4	1,73	0	
	23	6.75	2,25	6	3	1,30	0,75	

5	24	5.25	1,75	4	3	1,15	1,25	превышение
	25	4.5	1,5	3	3	1,06	1,5	превышение
	26	6	2	7	1	1,22	1	
	27	3.75	1,25	5	0	0,97	1,25	превышение
6	28	17.25	5,75	17	6	2,08	0,25	
	29	7.5	2,5	4	6	1,37	3,5	превышение
	30	12	4	12	4	1,73	0	
	31	8.25	2,75	8	3	1,44	0,25	
	32	14.25	4,75	15	4	1,89	0,75	
	33	9.75	3,25	8	5	1,56	1,75	превышение
7	34	5.25	1,75	5	2	1,15	0,25	
	35	6	2	5	3	1,22	1	
	36	12.75	4,25	12	5	1,79	0,75	
	37	5.25	1,75	6	1	1,15	0,75	
	38	23.25	7,75	18	13	2,41	5,25	превышение
	39	3.75	1,25	4	1	0,97	0,25	
	40	3.75	1,25	3	2	0,97	0,75	
	41	4.5	1,5	5	1	1,06	0,5	
	42	12	4	8	8	1,73	4	превышение
	43	9	3	8	4	1,50	1	
8	44	13.5	4,5	15	3	1,84	1,5	
	45	6.75	2,25	7	2	1,30	0,25	
	46	19.5	6,5	23	3	2,21	3,5	превышение
	47	9.75	3,25	12	1	1,56	2,25	превышение
	48	15.75	5,25	18	3	1,98	2,25	превышение
	49	8.25	2,75	11	0	1,44	2,75	превышение
	51	7.5	2,5	8	2	1,37	0,5	
	52	3.75	1,25	3	2	0,97	0,75	
	53	8.25	2,75	7	4	1,44	1,25	
9	54	12.75	4,25	14	3	1,79	1,25	
	55	18	6	17	7	2,12	1	
	56	4.5	1,5	4	2	1,06	0,5	
	57	15	5	16	4	1,94	1	
	58	12.75	4,25	14	3	1,79	1,25	
	59	6	2	7	1	1,22	1	
	60	7.5	2,5	9	1	1,37	1,5	превышение
	61	12.75	4,25	10	7	1,79	2,75	превышение
	62	13.5	4,5	12	6	1,84	1,5	
10	63	14.25	4,75	15	4	1,89	0,75	
	64	4.5	1,5	5	1	1,06	0,5	
	65	12	4	11	5	1,73	1	
	66	1.5	0,5	2	0	0,61	0,5	
	67	21.75	7,25	21	8	2,33	0,75	
	68	13.5	4,5	13	5	1,84	0,5	
	69	8.25	2,75	8	3	1,44	0,25	
	70	13.5	4,5	17	1	1,84	3,5	превышение
	71	12	4	13	3	1,73	1	
	72	8.25	2,75	9	2	1,44	0,75	
	73	4.5	1,5	5	1	1,06	0,5	
	74	10.5	3,5	10	4	1,62	0,5	
11	75	10.5	3,5	11	3	1,62	0,5	
	76	6.75	2,25	9	0	1,30	2,25	превышение
	77	12	4	13	3	1,73	1	
	78	9	3	7	5	1,50	2	превышение
	79	12.75	4,25	14	3	1,79	1,25	
	80	3	1	3	1	0,87	0	
	81	11.25	3,75	12	3	1,68	0,75	

	82	6.75	2,25	6	3	1,30	0,75	
12	83	9	3	8	4	1,50	1	
	84	12	4	12	4	1,73	0	
	85	10.5	3,5	9	5	1,62	1,5	
	86	5.25	1,75	5	2	1,15	0,25	
	88	2.25	0,75	2	1	0,75	0,25	
	89	5.25	1,75	4	3	1,15	1,25	превышение
	90	8.25	2,75	5	6	1,44	3,25	превышение
	91	9.75	3,25	9	4	1,56	0,75	
13	92	13.5	4,5	15	3	1,84	1,5	
	93	19.5	6,5	23	3	2,21	3,5	превышение
	94	6.75	2,25	8	1	1,30	1,25	
	95	6.75	2,25	8	1	1,30	1,25	
	96	11.25	3,75	13	2	1,68	1,75	превышение
	97	12.75	4,25	0	17	1,79	12,75	превышение
	98	7.5	2,5	10	0	1,37	2,5	превышение
	99	6.75	2,25	9	0	1,30	2,25	превышение
	100	6.75	2,25	7	2	1,30	0,25	

Таблица 6 из [10]

Расщепление в F_2 по окраске семядолей (по семьям).

№ комб.	№ семьи	Ожидаемое соотн. (3 : 1)		Полученное соотн.		Ошибка ±	Разница м-ду ожид. и получ.	превышение
1	22	11,25	3,75	11	4	1,68	0,25	
	23	26,25	8,75	27	8	2,56	0,75	
	24	12,75	4,25	12	5	1,79	0,75	
	25	15,75	5,25	14	7	1,98	1,75	
	26	4,5	1,5	5	1	1,06	0,5	
	27	18,75	6,25	20	5	2,17	1,25	
	28	40,5	13,5	43	11	3,18	2,5	
	29	21	7	25	3	2,29	4	превышение
2	30	19,5	6,5	16	10	2,21	3,5	превышение
	31	23,25	7,75	25	6	2,41	1,75	
	32	17,25	5,75	15	8	2,08	2,25	превышение
	33	23,25	7,75	26	5	2,41	2,75	превышение
	34	9,75	3,25	12	1	1,56	2,25	превышение
	35	9,75	3,25	11	2	1,56	1,25	
	36	20,25	6,75	15	12	2,25	5,25	превышение
3	37	19,5	6,5	20	6	2,21	0,5	
	38	6,75	2,25	5	4	1,30	1,75	превышение
	39	7,5	2,5	8	2	1,37	0,5	
	40	25,5	8,5	19	15	2,52	6,5	превышение
4	41	16,5	5,5	9	13	2,03	7,5	превышение
	42	11,25	3,75	10	5	1,68	1,25	
	43	21	7	20	8	2,29	1	
	44	14,25	4,75	15	4	1,89	0,75	
	45	19,5	6,5	18	8	2,21	1,5	
	46	24	8	27	5	2,45	3	превышение
5	47	14,25	4,75	14	5	1,89	0,25	
	48	9,75	3,25	8	5	1,56	1,75	превышение
	49	22,5	7,5	26	4	2,37	3,5	превышение
	50	12	4	11	5	1,73	1	
	51	7,5	2,5	4	6	1,37	3,5	превышение
	52	6,75	2,25	6	3	1,30	0,75	

	53	14,25	4,75	13	6	1,89	1,25	
	54	15	5	13	7	1,94	2	превышение
	55	11,25	3,75	10	5	1,68	1,25	
	56	14,25	4,75	11	8	1,89	3,25	превышение
	57	18	6	19	5	2,03	1	
6	58	16,5	5,5	17	5	1,68	0,5	
	59	11,25	3,75	11	4	2,41	0,25	
	60	23,25	7,75	21	10	2,03	2,25	
	61	16,5	5,5	18	4	1,44	1,5	
	62	8,25	2,75	8	3	1,98	0,25	
	63	15,75	5,25	16	5	1,94	0,25	
	64	15	5	17	3	1,62	2	превышение
	65	10,5	3,5	11	3	1,98	0,5	
7	66	15,75	5,25	16	5	2,74	0,25	
	67	30	10	31	9	2,12	1	
	68	18	6	18	6	1,15	0	
	69	5,25	1,75	5	2	1,98	0,25	
	70	15,75	5,25	15	6	1,89	0,75	
	71	14,25	4,75	15	4	1,98	0,75	
	72	15,75	5,25	16	5	1,98	0,25	
8	73	15,75	5,25	17	4	2,37	1,25	
	74	22,5	7,5	22	8	1,84	0,5	
	75	13,5	4,5	13	5	2,08	0,5	
	76	17,25	5,75	19	4	1,68	1,75	
	77	11,25	3,75	12	3	1,89	0,75	
	78	14,25	4,75	13	6	2,29	1,25	
	79	21	7	22	6	2,49	1	
	80	24,75	8,25	29	4	2,08	4,25	превышение
	81	17,25	5,75	16	7	1,68	1,25	
9	82	11,25	3,75	11	4	2,17	0,25	
	83	18,75	6,25	22	3	2,12	3,25	превышение
	84	18	6	20	4	2,08	2	
	85	17,25	5,75	17	6	2,37	0,25	
	86	22,5	7,5	21	9	1,89	1,5	
	87	14,25	4,75	13	6	1,56	1,25	
	88	9,75	3,25	8	5	3,27	1,75	превышение
	89	42,75	14,25	44	13	2,56	1,25	
	90	26,25	8,75	28	7	2,41	1,75	
	91	23,25	7,75	21	10	2,52	2,25	
10	93	25,5	8,5	27	7	2,33	1,5	
	94	21,75	7,25	22	7	0,00	0,25	
	96	18	6	17	7	2,12	1	
	97	16,5	5,5	13	9	2,03	3,5	превышение
	98	10,5	3,5	10	4	1,62	0,5	
	99	6,75	2,25	7	2	1,30	0,25	
	100	17,25	5,75	22	1	2,08	4,75	превышение
	101	27,75	9,25	25	12	2,63	2,75	превышение
	102	21	7	23	5	2,29	2	
11	103	6,75	2,25	6	3	1,30	0,75	
	104	14,25	4,75	16	3	1,89	1,75	
	105	37,5	12,5	50	0	3,06	12,5	превышение
12	106	15,75	5,25	15	6	1,98	0,75	
	107	24,75	8,25	26	7	2,49	1,25	
	108	14,25	4,75	14	5	1,89	0,25	
	109	10,5	3,5	9	5	1,62	1,5	
	110	22,5	7,5	22	8	2,37	0,5	
	111	15	5	14	6	1,94	1	

	112	15	5	12	8	1,94	3	превышение
	113	12,75	4,25	8	9	1,79	4,75	превышение
	114	21,75	7,25	23	6	2,33	1,25	
	116	13,5	4,5	12	6	1,84	1,5	
13	117	24,75	8,25	23	10	2,49	1,75	
	118	11,25	3,75	11	4	1,68	0,25	
	119	14,25	4,75	15	4	1,89	0,75	
	120	15	5	13	7	1,94	2	превышение
	121	23,25	7,75	24	7	2,41	0,75	
	122	12,75	4,25	15	2	1,79	2,25	превышение
	123	21,75	7,25	18	11	2,33	3,75	превышение
	124	15	5	13	7	1,94	2	превышение
	125	23,25	7,75	23	8	2,41	0,25	
14	126	17,25	5,75	20	3	2,08	2,75	превышение
	128	20,25	6,75	22	5	2,25	1,75	
	129	9	3	6	6	1,50	3	превышение
	130	19,5	6,5	23	3	2,21	3,5	превышение
	131	16,5	5,5	13	9	2,03	3,5	превышение
	132	17,25	5,75	17	6	2,08	0,25	
	133	12	4	12	4	1,73	0	
15	134	12	4	13	3	1,73	1	
	135	15	5	16	4	1,94	1	
	136	31,5	10,5	30	12	2,81	1,5	
	137	33	11	31	13	2,87	2	
	138	19,5	6,5	24	2	2,21	4,5	превышение
	139	18	6	19	5	2,12	1	
	140	6	2	3	5	1,22	3	превышение
	141	38,25	12,75	37	14	3,09	1,25	
	142	48	16	46	18	3,46	2	
16	143	19,5	6,5	19	7	2,21	0,5	
	145	10,5	3,5	7	7	1,62	3,5	превышение
	146	9	3	10	2	1,50	1	
	147	26,25	8,75	22	13	2,56	4,25	превышение
	148	7,5	2,5	0	10	1,37	7,5	превышение

Более ранняя работа Ермолаевой [11] касается того же вопроса о расщеплении признаков у гороха.

В этой работе сорта гороха, подлежащие скрещиванию, высевали в разные сроки, и также в разные сроки производилось скрещивание для получения гибридных семян F_1. На следующий год эти семена высевались и получались семена второго поколения F_2, для которых наблюдалось расщепление по окраске семядолей (желтая/зеленая). Надо сказать, что в этих опытах Ермолаева вообще не добилась менделевского явления доминирования. При скрещивании красноцветкового сорта гороха с белоцветковым семена поколения F_1 должны дать только красноцветковые растения. А Ермолаева наблюдала на 192 красноцветковых растения 32 белоцветковых. Расщепление признаков по окраске семядолей в потомстве этих 192 и 32 растений шло совершенно по разному. Среди семян красноцветковых растений несколько преобладали желтые (впрочем далеко не дотягивая до менделевского соотношения 3:1), а среди семян белоцветковых вообще большинство было зеленых. При этом никакой статистической устойчивостью соотношения между количеством желтых и зеленых семян не наблюдалось. Например, в первых двух строчках таблицы 1 растения первого срока посева скрещиваются соответственно, 13.08.1936 и 15.08.1936. В первом случае красноцветковые растения поколения F_1 дали 150 желтых и 139 зеленых семян, а во втором случае, соответственно 134 и 62. Можно ли принять гипотезу $p_1=p_2=p$ для двух серий испытаний Бернулли: $m_1=150$, $n_1=289$ и $m_2=134$, $n_2=2196$?. Обращаясь к пакету Statistica, находим, что соответствующее p-значение есть примерно $3\cdot10^{-4}$, т.е. гипотезу принять нельзя.

Есть в этой таблице 1 и также крайне маловероятные расщепления: 76:0, 131:0, а также 2:89. Все это исключает мысль о статистической однородности экспериментов и об их пригодности для оценки вероятности появления интересующего нас признака.. Сводные результаты в виде эмпирических функций и и ожидаемой функцией Лапласа (аналогично рис. 1 и 2) (отдельно для расщепления семян красноцветковых и белоцветковых растений) приводятся на рис. 8 и 9. Изображенные на этих рисунках эмпирические функции распределения не имеют ничего общего с нормальной кривой.

В таблице 2 этой работы приведены только выборочные результаты скрещиваний других гибридных комбинаций. Выборочность результатов делает бессмысленной их обработку. Приводим таблицу 1 из [11].

№№ комбинации по журналу	Комбинации	Дата скрещивания 1936г.	Окраска пазух листьев и цветка у 1-го поколения гибридов(F_1)			Расщепление F_2 гибридов (семена растений F_1)						
				из них		От красноцветковых растений			От белоцветковых растений			Отношение желто семядольных к зелено семядольным по вариантам
			Всего растений	Красноцветковых	Белоцветковых	Окраска семядолей		отношение	Окраска семядолей		отношение	
						желтая	зеленая		желтая	зеленая		
80	В.з.I х ГI *)	13.08	21	16	5	150	139	1,08	21	76	3,62	0,80
81	В.з.I х ГI	15.08	13	8	5	134	62	2,16	76	156	2,05	0,96
82	В.з.I х ГI	19.08	7	4	3	41	54	0,76	0	73		0,32
83	В.з.I х ГI	20.08	5	5	0	88	34	2,60	0	0		2,60
84	В.з.I х ГI	30.08	6	3	3	59	22	2,70	40	33	0,82	1,80
85	В.з.I х ГI	5.09	5	5	0	86	27	3,19	0	0		3,19
88	В.з.I х ГI	23.08	9	9	0	163	57	2,86	0	0		2,86
90	В.з.II х ГI	31.08	17	17	0	261	97	2,70	0	0		2,70
93	В.з.II х ГII	30.08	10	6	4	79	39	2,00	0	32		1,11
94	В.з.II х ГII	5.09	5	5	0	121	51	2,24***)	0	0		2,24***)
95	В.з.IV х ГII	8.09	4	4	0	83	28	2,96	0	0		2,96
98	В.з.I х В.зI **)	11.08	5	5	0	132	49	2,70	0	0		2,70
99	ГI х В.зI	15.08	4	4	0	90	44	2,04	0	0		2,04
100	ГI х В.зI	15.08	11	4(11?)	0***)	261	87	3,00	0	0		3,00
101	ГI х В.зI	31.08	12	12	0	247	98	2,52	0	0		2,52
102	ГI х В.зII	20.08	12	9	3	146	64	2,28	29	65	2,24	1,35
105	ГI х В.зIII	5.09	5	5	0	76	0		0	0		
104	ГI х В.зIII	30.08	11	11	0	240	84	2,86	0	0		2,86
106	ГII х В.зII	23.08	6	3	3	2	89	0,02	30	44	1,47	0,24

107	ГII х В.зII	28.08	8	8	0	241	68	3,54	0	0		3,54
108	ГII х В.зII	30.08	10	7	3	124	91	1,36	37	27	0,73	1,36
109	ГII х В.зII	5.09	8	7	1	121	87	1,40	0	18		1,15
110	ГIII х В.зIII	4.09	6	4	2	131	0		0	19		6,89
111	ГII х В.зIV	14.09	3	3	0	25	17	1,47	0	0		1,47
	"гигант"-контроль		150	150								
	"В.З."-контроль		70		70							

*) Сокращения: В.з.- сорт "Виктория зеленая", Г - "Гигант".
Римские цифры обозначают сроки посева родительских пар для скрещивания:
I - 27.06.1936; II - 9.07.1936; III - 20.07.1936; IV - 1.08.1936.
**) По-видимому, ошибка Ермолаевой: скрещивание сорта с самим собой бессмысленно.

***) По-видимому, ошибка автора.

Приложение 2. Данные Т.К.Енина

В работе [12] речь идет о расщеплении в поколении F_2 у томатов по признаку формы листьев (доминантный признак - нормально рассеченные листья, рецессивный признак – картофелевидные листья). Ниже приводится таблица с результатами.

Таблица из [12]. **Расщепление в F_2 от скрещивания томатов "Микадо" и "Эрлиана" (по семьям).**

№ раст ен. F_1	№ семьи F_2	Численн. семьи F_2	Из них с листьями		Теоретически с листьями		Ошибка + --	Разница между ожидаем ым и получен ным	превышение
			нормально рассеченны ми	картофеле- видными	нормально рассеченны ми	картофеле- видными			
5	1	365	270	95	273,75	91,25	8,27	3,75	
6	2	351	283	68	263,25	87,75	8,11	19,75	превышение
7	3	100	78	22	75	25	4,33	3	
8	4	497	381	116	372,75	124,25	9,65	8,25	
14	5	218	164	54	163,5	54,5	6,39	0,5	
16	6	259	195	64	194,25	64,75	6,97	0,75	
17	7	143	104	39	107,25	35,75	5,18	3,25	
18	8	319	252	67	239,25	79,75	7,73	12,75	превышение
21	9	211	161	50	158,25	52,75	6,29	2,75	
32	10	215	171	44	161,25	53,75	6,35	9,75	превышение
2	11	482	365	117	361,5	120,5	9,51	3,5	
12	12	395	294	101	296,25	98,75	8,61	2,25	
11	13	471	363	108	353,25	117,75	9,40	9,75	превышение
13	14	257	196	61	192,75	64,25	6,94	3,25	
19	15	692	518	174	519	173	11,39	1	
22	16	534	400	134	400,5	133,5	10,01	0,5	
23	17	595	448	147	446,25	148,75	10,56	1,75	
26	18	266	205	61	199,5	66,5	7,06	5,5	
27	19	694	519	175	520,5	173,5	11,41	1,5	

34	20	447	330	117	335,25	111,75	9,15	5,25	
35	21	353	265	88	264,75	88,25	8,14	0,25	
36	22	234	175	59	175,5	58,5	6,62	0,5	
38	23	438	333	105	328,5	109,5	9,06	4,5	
39	24	219	166	53	164,25	54,75	6,41	1,75	
40	25	416	315	101	312	104	8,83	3	
Сумма		9171	6951	2220	6878,25	2292,75			

В работе [13] рассматриваются результаты скрещивания различных сортов гороха по «менделирующим» признакам. В работе приведены четыре таблицы результатов (очевидные ошибки и опечатки мы устранили) Из них таблицы 1,2 и 3 относятся к одним и тем же скрещиваниям, но к различным менделирующим признакам. Номера растений F_1 (т.е. номера семей) имеют пропуски, но в данном случае автор указывает корректный принцип отбора: для таблиц 1 и 2 требуется, чтобы общее число семян поколения $F_{2,}$ полученных от данного растения F_1 было не меньше 20. Для таблицы 3 требуется, чтобы, из этих семян было получено не менее 20 растений, для которых можно наблюдать форму боба (часть семян может не взойти).Таблица 4 относится к другим скрещиваниям.

Таблица 1. Расщепление F_2 от скрещивания «Гороха Жегалова» на «Бисмарк» по окраске семядолей (семена растений F_1) (теоретическое отношение 3:1)

№ раст F_1	Эмпирич. число		Теоретич. число		ошибка
	желт.	зелен.	желт.	зелен.	
1	15	6	15,75	5,25	1,98
2	22	4	19,5	6,5	2,21
5	13	9	16,5	5,5	2,03
12	18	5	17,25	5,75	2,08
14	22	10	24	8	2,45
15	24	6	22,5	7,5	2,37
17	30	14	33	11	2,87
18	20	4	18	6	2,12
19	15	8	17,25	5,75	2,08
22	23	8	23,25	7,75	2,41
23	22	11	24,75	8,25	2,49
24	19	7	19,5	6,5	2,21
26	19	8	20,25	6,75	2,25
27	14	6	15	5	1,94
28	17	6	17,25	5,75	2,08
29	17	5	16,5	5,5	2,03
30	22	4	19,5	6,5	2,21
31	16	12	21	7	2,29
32	19	7	19,5	6,5	2,21
33	24	8	24	8	2,45
35	23	4	20,25	6,75	2,25
36	31	8	29,25	9,75	2,70
39	20	9	21,75	7,25	2,33
43	23	10	24,75	8,25	2,49
45	39	12	38,25	12,75	3,09
48	14	5	14,25	4,75	1,89
54	21	7	21	7	2,29
55	25	9	25,5	8,5	2,52
56	17	6	17,25	5,75	2,08
57	22	9	23,25	7,75	2,41
59	17	4	15,75	5,25	1,98
61	21	4	18,75	6,25	2,17
63	38	17	41,25	13,75	3,21
67	18	7	18,75	6,25	2,17
69	14	5	14,25	4,75	1,89

69	22	9	23,25	7,75	2,41
70	20	3	17,25	5,75	2,08
80	21	9	22,5	7,5	2,37
82	18	8	19,5	6,5	2,21
84	33	12	33,75	11,25	2,90
85	14	5	14,25	4,75	1,89
87	31	7	28,5	9,5	2,67
88	20	6	19,5	6,5	2,21
89	35	12	35,25	11,75	2,97
90	18	8	19,5	6,5	2,21
91	20	8	21	7	2,29
93	18	7	18,75	6,25	2,17
94	22	8	22,5	7,5	2,37
96	21	8	21,75	7,25	2,33
98	20	5	18,75	6,25	2,17
103	18	5	17,25	5,75	2,08
105	18	6	18	6	2,12
104	25	6	23,25	7,75	2,41
107	18	6	18	6	2,12
108	21	8	21,75	7,25	2,33
110	15	5	15	5	1,94
112	18	7	18,75	6,25	2,17
113	34	13	35,25	11,75	2,97
114	18	9	20,25	6,75	2,25
115	30	10	30	10	2,74
116	17	8	18,75	6,25	2,17
117	26	7	24,75	8,25	2,49
119	40	12	39	13	3,12
120	30	8	28,5	9,5	2,67
122	20	7	20,25	6,75	2,25
125	22	6	21	7	2,29
126	19	5	18	6	2,12
127	29	8	27,75	9,25	2,63
128	15	7	16,5	5,5	2,03
129	20	3	17,25	5,75	2,08
130	18	5	17,25	5,75	2,08
131	15	7	16,5	5,5	2,03
132	27	8	26,25	8,75	2,56
134	23	11	25,5	8,5	2,52
136	15	6	15,75	5,25	1,98
137	20	6	19,5	6,5	2,21
139	16	5	15,75	5,25	1,98
142	28	10	28,5	9,5	2,67
145	15	7	16,5	5,5	2,03
146	15	9	18	6	2,12
147	16	4	15	5	1,94
148	17	5	16,5	5,5	2,03
149	15	7	16,5	5,5	2,03
150	13	7	15	5	1,94
153	21	6	20,25	6,75	2,25
155	20	7	20,25	6,75	2,25
156	28	10	28,5	9,5	2,67
157	21	9	22,5	7,5	2,37
158	14	6	15	5	1,94

159	22	7	21,75	7,25	2,33
160	20	7	20,25	6,75	2,25
165	25	3	21	7	2,29
166	16	5	15,75	5,25	1,98
169	22	7	21,75	7,25	2,33
170	16	5	15,75	5,25	1,98
171	35	12	35,25	11,75	2,97
172	23	8	23,25	7,75	2,41
173	16	4	15	5	1,94
174	25	11	27	9	2,60
177	33	9	31,5	10,5	2,81
180	17	8	18,75	6,25	2,17
182	27	5	24	8	2,45
184	19	11	22,5	7,5	2,37
185	14	7	15,75	5,25	1,98
186	19	6	18,75	6,25	2,17
187	27	6	24,75	8,25	2,49
188	14	6	15	5	1,94
197	21	6	20,25	6,75	2,25
200	14	6	15	5	1,94
201	16	4	15	5	1,94
202	22	6	21	7	2,29
203	27	7	25,5	8,5	2,52
204	22	9	23,25	7,75	2,41
итого	2389	825	2410,5	803,5	24,55

Таблица2. Расщепление F_2 от скрещивания «Гороха Жегалова» на «Бисмарк» по форме семян (семена растений F_1) (теоретическое отношение 3:1)

№ раст F_1	Эмпирич. число		Теоретич. число		ошибка
	кругл	морщин	кругл	морщин	
1	16	5	15,75	5,25	1,98
2	19	7	19,5	6,5	2,21
5	17	5	16,5	5,5	2,03
12	17	6	17,25	5,75	2,08
14	24	8	24	8	2,45
15	24	6	22,5	7,5	2,37
17	35	9	33	11	2,87
18	15	9	18	6	2,12
19	18	5	17,25	5,75	2,08
22	24	7	23,25	7,75	2,41
23	25	8	24,75	8,25	2,49
24	19	7	19,5	6,5	2,21
26	20	7	20,25	6,75	2,25
27	16	4	15	5	1,94
28	19	4	17,25	5,75	2,08
29	19	3	16,5	5,5	2,03
30	19	7	19,5	6,5	2,21
31	22	6	21	7	2,29
32	20	6	19,5	6,5	2,21
33	29	3	24	8	2,45
35	25	2	20,25	6,75	2,25
36	30	9	29,25	9,75	2,70
39	23	6	21,75	7,25	2,33

43	29	4	24,75	8,25	2,49
45	41	11	39	13	3,12
48	14	5	14,25	4,75	1,89
54	21	7	21	7	2,29
55	26	8	25,5	8,5	2,52
56	18	5	17,25	5,75	2,08
57	21	10	23,25	7,75	2,41
59	16	5	15,75	5,25	1,98
61	19	6	18,75	6,25	2,17
63	31	14	33,75	11,25	2,90
67	20	5	18,75	6,25	2,17
68	15	4	14,25	4,75	1,89
69	23	8	23,25	7,75	2,41
70	20	3	17,25	5,75	2,08
80	23	7	22,5	7,5	2,37
82	19	7	19,5	6,5	2,21
84	37	8	33,75	11,25	2,90
85	15	4	14,25	4,75	1,89
87	29	9	28,5	9,5	2,67
88	19	7	19,5	6,5	2,21
89	33	13	34,5	11,5	2,94
90	18	8	19,5	6,5	2,21
91	20	8	21	7	2,29
93	18	7	18,75	6,25	2,17
94	21	9	22,5	7,5	2,37
96	22	7	21,75	7,25	2,33
98	17	8	18,75	6,25	2,17
103	17	6	17,25	5,75	2,08
104	24	7	23,25	7,75	2,41
105	18	6	18	6	2,12
107	18	6	18	6	2,12
108	22	7	21,75	7,25	2,33
110	17	3	15	5	1,94
112	18	7	18,75	6,25	2,17
113	34	13	35,25	11,75	2,97
114	19	8	20,25	6,75	2,25
115	30	10	30	10	2,74
116	24	1	18,75	6,25	2,17
117	26	7	24,75	8,25	2,49
119	39	13	39	13	3,12
120	29	9	28,5	9,5	2,67
122	18	9	20,25	6,75	2,25
125	20	8	21	7	2,29
126	18	6	18	6	2,12
127	28	9	27,75	9,25	2,63
128	17	5	16,5	5,5	2,03
129	16	7	17,25	5,75	2,08
130	18	5	17,25	5,75	2,08
131	17	5	16,5	5,5	2,03
132	27	8	26,25	8,75	2,56
134	25	9	25,5	8,5	2,52
136	15	6	15,75	5,25	1,98
137	22	4	19,5	6,5	2,21
139	16	5	15,75	5,25	1,98

142	32	6	28,5	9,5	2,67
145	17	5	16,5	5,5	2,03
146	19	5	18	6	2,12
147	17	3	15	5	1,94
148	17	5	16,5	5,5	2,03
149	17	5	16,5	5,5	2,03
150	16	4	15	5	1,94
153	22	5	20,25	6,75	2,25
155	21	6	20,25	6,75	2,25
156	29	9	28,5	9,5	2,67
157	26	4	22,5	7,5	2,37
158	17	3	15	5	1,94
159	22	7	21,75	7,25	2,33
160	24	3	20,25	6,75	2,25
165	23	5	21	7	2,29
166	16	5	15,75	5,25	1,98
169	24	5	21,75	7,25	2,33
170	18	3	15,75	5,25	1,98
171	38	9	35,25	11,75	2,97
172	23	8	23,25	7,75	2,41
173	17	5	16,5	5,5	2,03
174	28	8	27	9	2,60
177	30	12	31,5	10,5	2,81
180	19	6	18,75	6,25	2,17
182	24	8	24	8	2,45
184	23	7	22,5	7,5	2,37
185	16	5	15,75	5,25	1,98
186	23	3	19,5	6,5	2,21
187	26	7	24,75	8,25	2,49
188	17	3	15	5	1,94
197	20	7	20,25	6,75	2,25
200	15	5	15	5	1,94
201	16	4	15	5	1,94
202	21	7	21	7	2,29
203	25	9	25,5	8,5	2,52
204	24	7	23,25	7,75	2,41
итого	2474	733	2405,25	801,75	

Таблица 3. Расщепление F_2 от скрещивания «Гороха Жегалова» на «Бисмарк» по форме боба (теоретическое отношение 9:7)

№ растения F_1	Эмпирическое число		Теоретическое число		ошибка
	лущ.	сахар.	лущ.	сахар.	
2	15	9	13,50	10,50	2,43
5	13	7	11,25	8,75	2,22
12	15	7	12,38	9,63	2,33
14	20	12	18,00	14,00	2,81
15	18	11	16,31	12,69	2,67
17	31	11	23,63	18,38	3,21
18	15	8	12,94	10,06	2,38
24	19	12	17,44	13,56	2,76
22	15	10	14,06	10,94	2,48
23	14	9	12,94	10,06	2,38
25	13	7	11,25	8,75	2,22
27	13	8	11,81	9,19	2,27

28	13	9	12,38	9,63	2,33
29	14	10	13,50	10,50	2,43
30	17	11	15,75	12,25	2,63
35	15	11	14,63	11,38	2,53
39	13	10	12,94	10,06	2,38
43	15	11	14,63	11,38	2,53
45	25	19	24,75	19,25	3,29
54	14	8	12,38	9,63	2,33
55	19	13	18,00	14,00	2,81
61	12	10	12,38	9,63	2,33
63	24	17	23,06	17,94	3,18
70	11	9	11,25	8,75	2,22
84	20	15	19,69	15,31	2,93
87	19	11	16,88	13,13	2,72
88	13	9	12,38	9,63	2,33
89	26	18	24,75	19,25	3,29
90	15	7	12,38	9,63	2,33
91	14	11	14,06	10,94	2,48
93	12	10	12,38	9,63	2,33
94	14	9	12,94	10,06	2,38
96	15	10	14,06	10,94	2,48
98	13	10	12,94	10,06	2,38
105	13	10	12,94	10,06	2,38
107	17	6	12,94	10,06	2,38
108	14	10	13,50	10,50	2,43
112	16	9	14,06	10,94	2,48
113	19	12	17,44	13,56	2,76
115	14	9	12,94	10,06	2,38
116	15	11	14,63	11,38	2,53
117	16	14	16,88	13,13	2,72
119	23	14	20,81	16,19	3,02
120	18	8	14,63	11,38	2,53
122	14	10	13,50	10,50	2,43
129	18	9	15,19	11,81	2,58
130	11	9	11,25	8,75	2,22
131	13	10	12,94	10,06	2,38
132	18	13	17,44	13,56	2,76
134	13	11	13,50	10,50	2,43
136	12	8	11,25	8,75	2,22
142	21	14	19,69	15,31	2,93
146	14	10	13,50	10,50	2,43
148	13	9	12,38	9,63	2,33
155	12	8	11,25	8,75	2,22
156	11	9	11,25	8,75	2,22
157	15	5	11,25	8,75	2,22
160	12	8	11,25	8,75	2,22
165	11	9	11,25	8,75	2,22
169	14	9	12,94	10,06	2,38
171	19	9	15,75	12,25	2,63
174	15	5	11,25	8,75	2,22
177	14	8	12,38	9,63	2,33
180	11	9	11,25	8,75	2,22
187	12	8	11,25	8,75	2,22
итого	1012	652	936	728	

Таблица 4. Расщепление F_2 от скрещивания «Чуда Лиона» на «Новую линию» по высоте стебля и окраске цветка (теоретическое отношение 3:1)

№ рас тения F_1	Эмпирическое число		Теоретическое число		Эмпирическое число		Теоретическое число		ошибка
	высоких	карликов.	высоких	карликов.	окраш	неокраш	окраш	неокраш	
1	17	7	18,00	6,00	19	5	18	6	2,12
2	23	8	23,25	7,75	23	8	23,25	7,75	2,41
3	16	5	15,75	5,25	16	5	15,75	5,25	1,98
4	18	5	17,25	5,75	18	6	18	6	2,12
5	13	7	15,00	5,00	14	8	16,5	5,5	2,03
6	19	6	18,75	6,25	19	6	18,75	6,25	2,17
7	13	6	14,25	4,75	15	4	14,25	4,75	1,89
8	16	7	17,25	5,75	17	6	17,25	5,75	2,08
9	12	4	12,00	4,00	12	4	12	4	1,73
10	12	6	13,50	4,50	14	4	13,5	4,5	1,84
11	16	5	15,75	5,25	14	7	15,75	5,25	1,98
12	19	5	18,00	6,00	18	7	18,75	6,25	2,17
13	30	12	31,50	10,50	35	7	31,5	10,5	2,81
14	15	6	15,75	5,25	19	2	15,75	5,25	1,98
15	22	4	19,50	6,50	20	6	19,5	6,5	2,21

Для каждой таблицы были построены графики эмпирической функции распределения величин $\{x_i\}$ в сравнении с теоретической $N(0,1)$ в обычном и нормальном масштабе (рис11-14), оценены их средние и стандартные отклонения и вычислены значения статистики критерия Колмогорова.

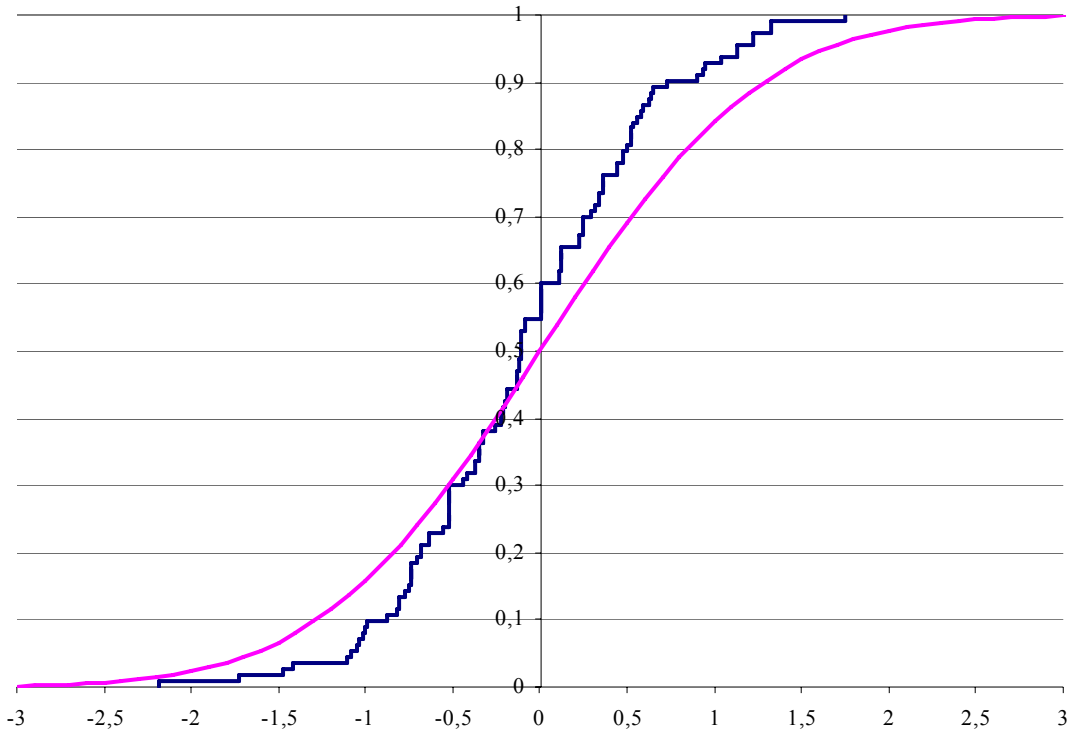

Рис.11*а* . Данные Енина (табл.1 из [13]) в обычном масштабе

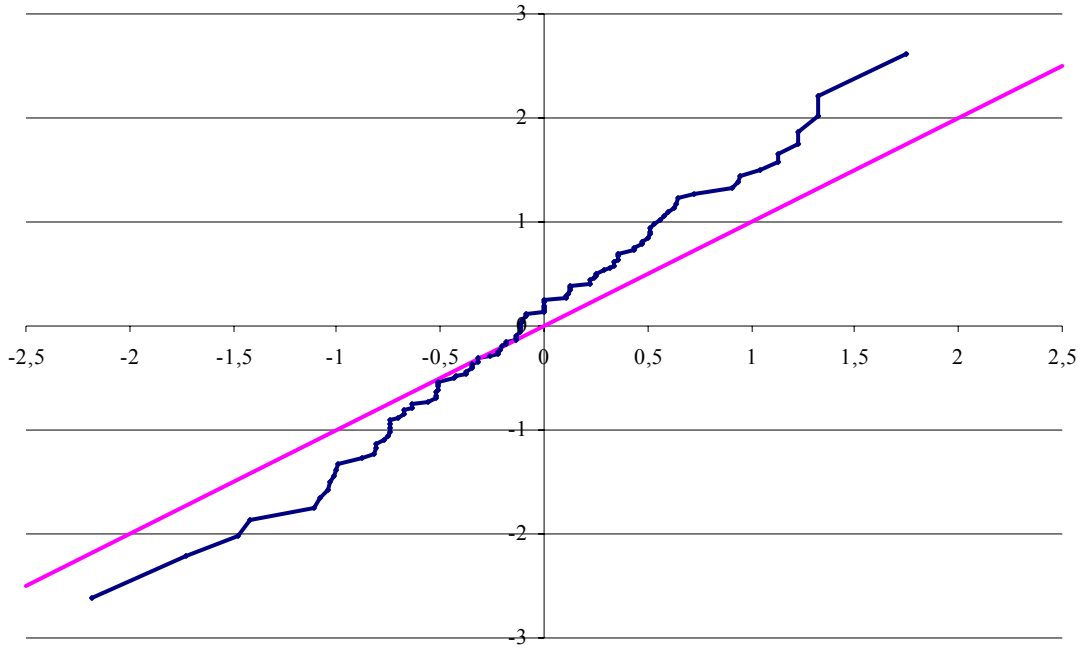

Рис.11*б* . Данные Енина (табл.1 из [13]) в нормальном масштабе.

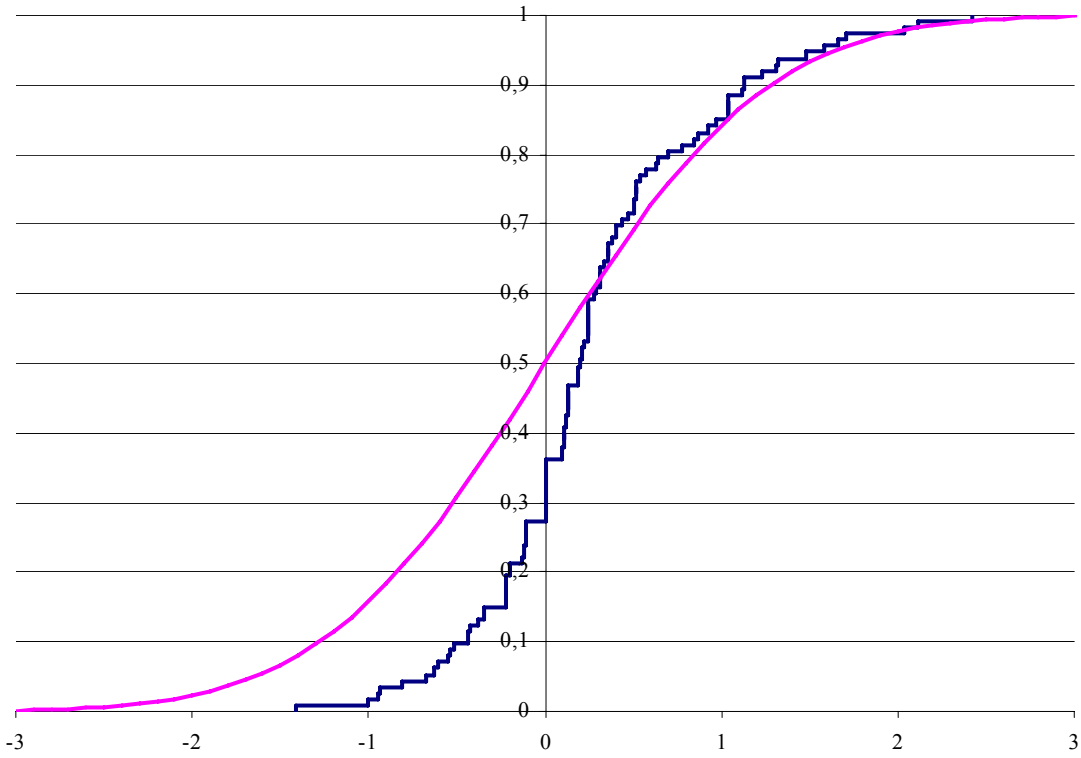

Рис.12*а*. Данные Енина (табл.2 из [13]) в обычном масштабе.

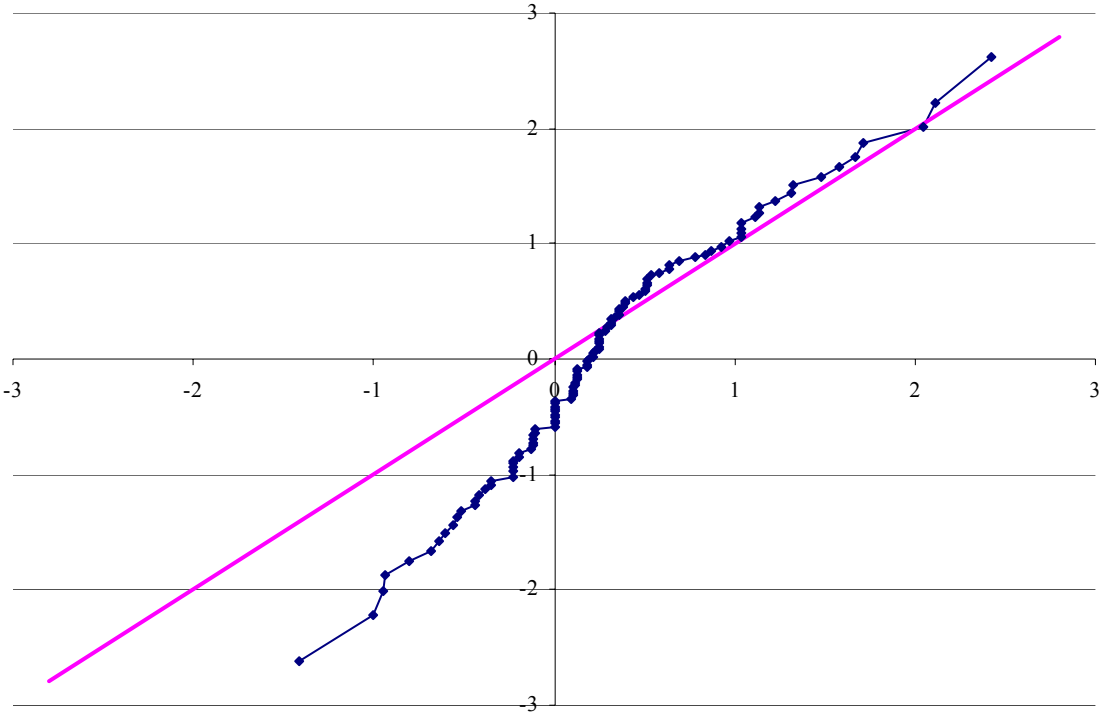

Рис.12*б* . Данные Енина (табл.2 из [13]) в нормальном масштабе.

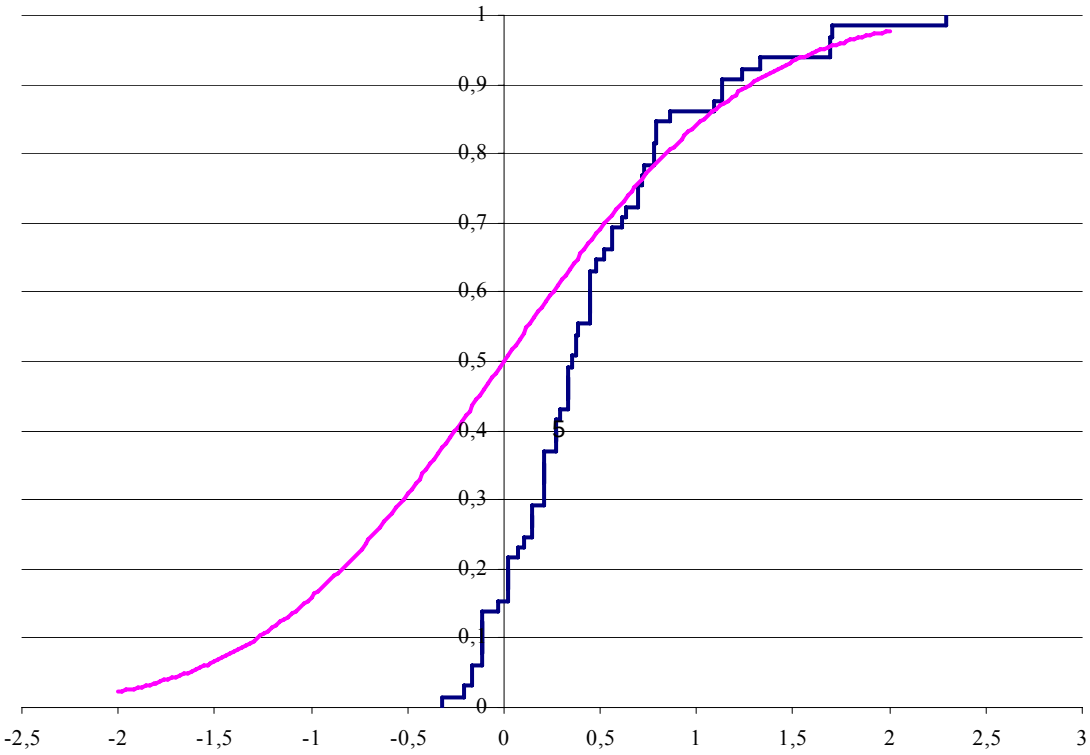

Рис.13*а* . Данные Енина (табл.2 из [13]) в обычном масштабе.

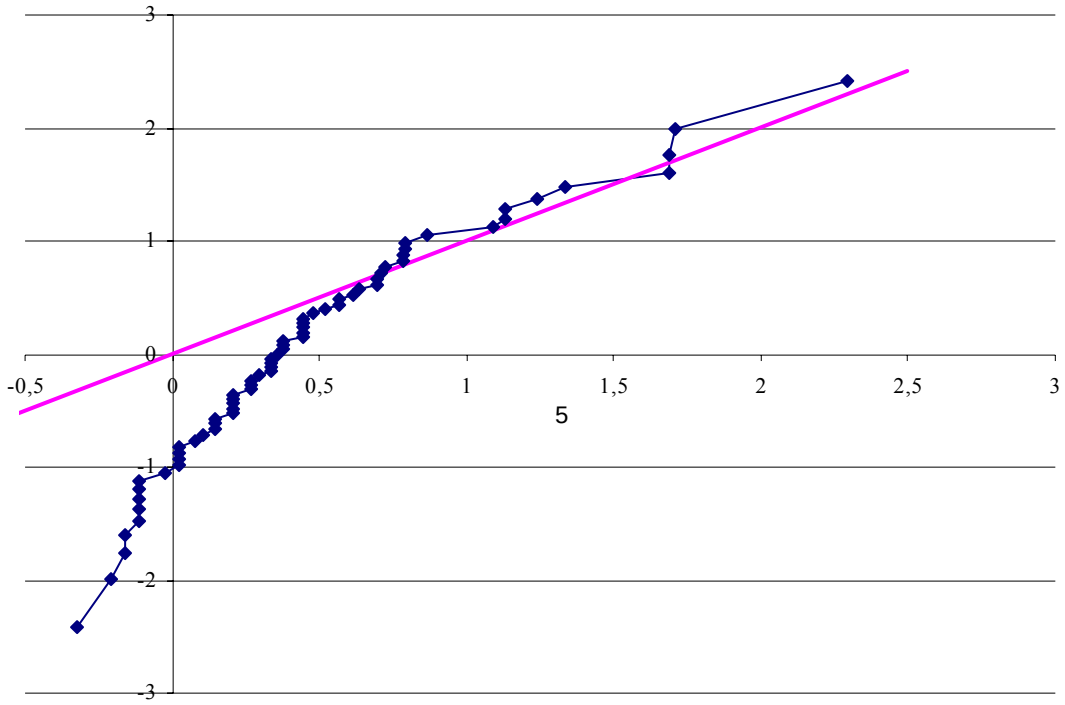

Рис.13*б* . Данные Енина (табл.3 из [13]) в нормальном масштабе.

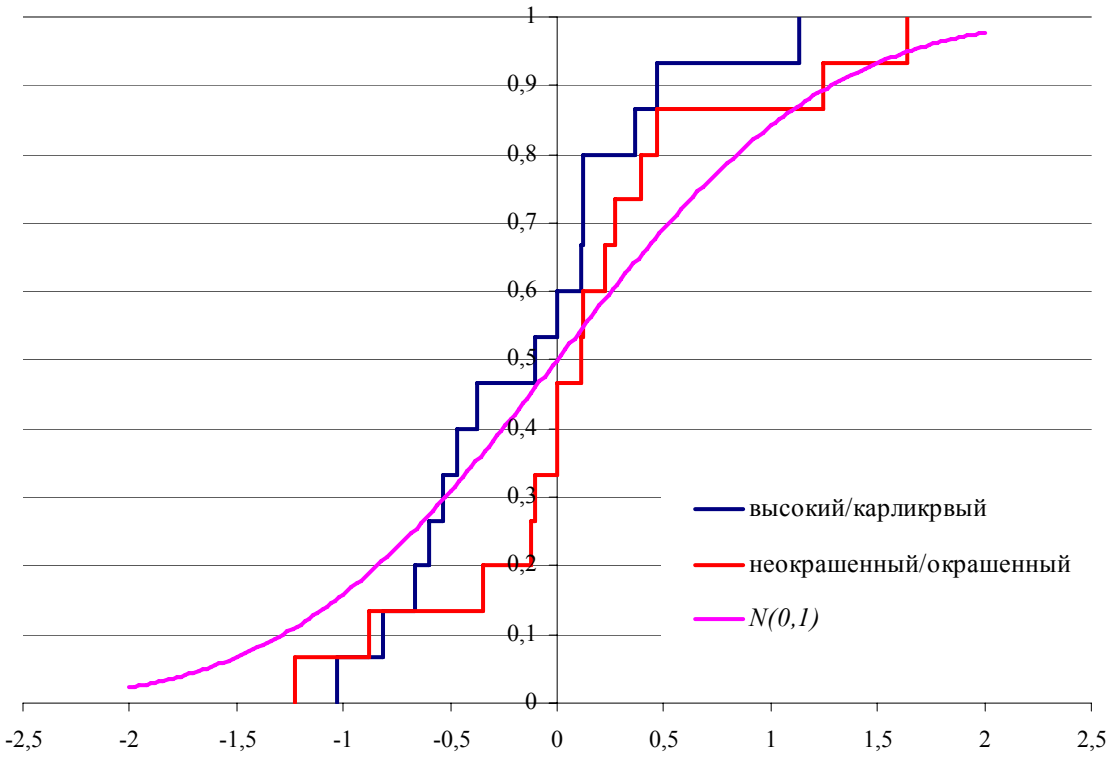

Рис.14*а* . Данные Енина (табл. 4 из [13]) в обычном масштабе.

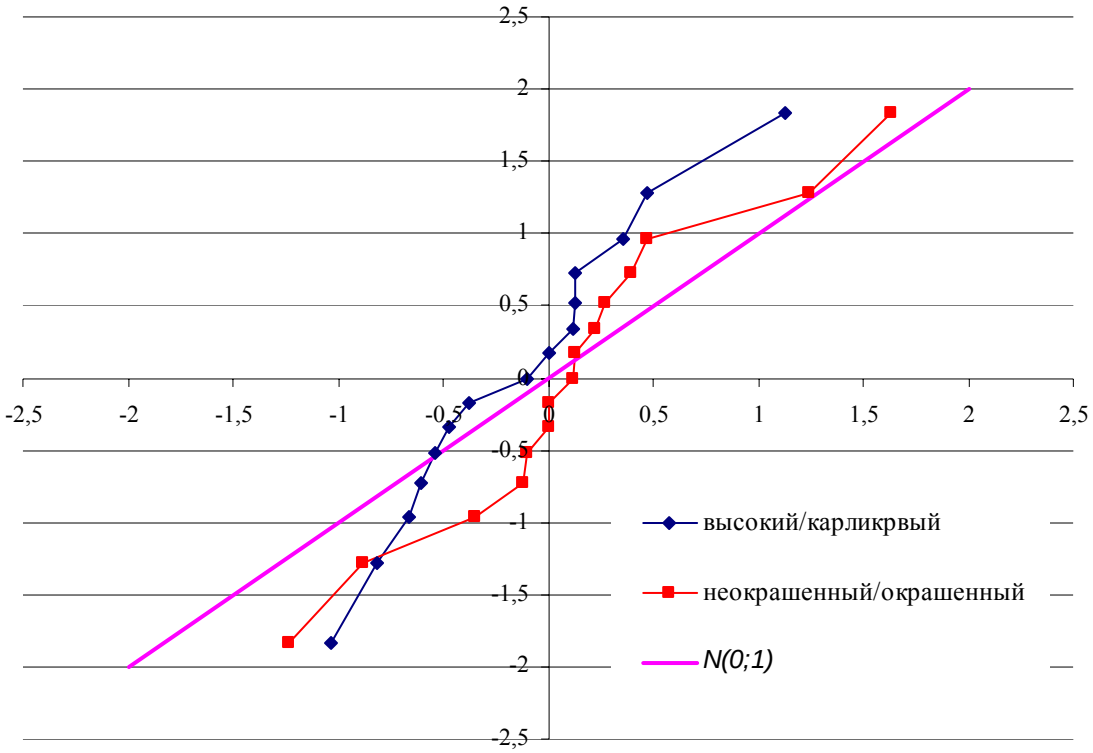

Рис.14*б* . Данные Енина (табл. 4 из [13]) в нормальном масштабе.

Результаты таковы:

№ таб	*r*-число семей	$\bar{x}$	s_x	*D*	λ	*p*
1	113	-0,084	0,6830	0,153	1,622	0,010
2	113	0,271	0,6539	0,260	2,764	4,65E-07
3	65	0,462	0,5168	0,405	3,267	1,07E-09
4 (по высоте)	15	0,121	0,7067	0,252	0,976	0,296
4 (по окраске)	15	-0,151	0,5690	0,252	0,9760	0,296

Как видно из графиков и из таблицы и эти данные имеют тенденцию к уменьшению (по сравнению с ожидаемым σ=1) стандартного отклонения, как и в [12]. Для таблиц 2 и 3 , кроме того, среднее значение $\bar{x}$ высоко значимо отклоняется от нуля.