Введение

При выполнении курсового проекта по математической статистике возникает много вопросов как по поводу теоретического обоснования применяемых процедур, так и по поводу их практической реализации. В данном пособии даются описания теоретических основ применения этих процедур. Схемы вычислений в рамках популярного компьютерного приложения MS Excel приведены в пособии [4]. По соображениям полноты картины, к сожалению, пришлось расширить до 16 общее число заданий – приблизительно по одному на каждую неделю семестра.

Работу над курсовым проектом следует начать с изучения главы “Предварительные понятия и определения”. Эта глава будет весьма полезна при подготовке ответов на контрольные вопросы. Выполнение каждого задания лучше всего начинать с изучения теоретического обоснования тех процедур, которые рассматриваются в этом задании. Причем желательно изучить весь материал заранее, до проведения соответствующего занятия в компьютерном классе.

Задания

Задание 1. Вычислить выборочные характеристики – среднее, дисперсию, стандартное отклонение, асимметрию, эксцесс.

Задание 2. Построить гистограмму выборки с подогнанной нормальной (равномерной, экспоненциальной) плотностью.

Задание 3. Построить эмпирическую функцию распределения выборки с подогнанной нормальной (равномерной, экспоненциальной) функцией распределения.

Задание 4. Проверить гипотезу нормальности (равномерности, экспоненциальности) выборочных данных.

Задание 5. Проверить гипотезу однородности по одновыборочному критерию Стьюдента.

Задание 6. Проверить гипотезу однородности по критерию знаков.

Задание 7. Проверить гипотезу однородности по двухвыборочному критерию Стьюдента.

Задание 8. Проверить гипотезу однородности по критерию Вилкоксона.

Задание 9. Проверить гипотезу равенства дисперсий.

Задание 10. Проверить гипотезу однородности по критерию хи-квадрат.

Задание 11. Построить доверительные границы для среднего значения нормального распределения.

Задание 12. Построить доверительные границы для дисперсии нормального распределения.

Задание 13. Построить доверительные границы для вероятности успеха.

Задание 14. Проверить гипотезу независимости признаков по критерию сопряженности хи-квадрат.

Задание 15. Проверить гипотезу независимости по критерию Стьюдента.

Задание 16. Построить линии регрессии.

Глава I. Предварительные понятия и определения

Выполнение курсового проекта по математической статистике требует от студента знания некоторых основ теории статистического вывода. В этой главе в краткой форме будет дано изложение этих основ.

§ 1. *Выборка*

Предположим, что в эксперименте наблюдается реализация x некоторой случайной величины (с.в.) X. Распределение этой с.в.

$$F(x) = \mathbf{P}\{X < x\}$$

неизвестно или известно с точностью до некоторого (возможно, векторного) параметра θ: $F(x) = F_\theta(x) = F(x|\theta)$. *Выборкой* объема n называется вектор $x^{(n)} = (x_1,...,x_n)$ независимых реализаций с.в. X. Более точно следует говорить о реализации независимых одинаково распределенных с.в. $X_1,...,X_n$. В связи с этим возникает возможность вычисления вероятностей тех или иных событий, связанных с выборкой. Тот факт, что эта вероятность (или соответствующие вероятностные характеристики) вычисляется при истинном значении параметра, равном θ, будет обозначаться значком θ у символов вероятности $\mathbf{P}_\theta$, мат.ожидания $\mathbf{E}_\theta$, дисперсии $\mathbf{D}_\theta$, … .

Задача статистического анализа состоит в принятии решений относительно распределения $F_\theta(x)$ наблюдаемой в эксперименте с.в. X. Чаще всего эта задача формулируется в терминах неизвестного значения параметра θ. Решение обычно принимается на основе некоторой *статистики* $T = T(x^{(n)})$ – функции выборочных данных, принимающей значения в пространстве возможных решений и не зависящей от неизвестных параметров вероятностной модели. Рассмотрим наиболее популярные статистические задачи – проверку гипотез и оценивание.

§ 2. *Проверка гипотез*

♦ **Гипотеза.**

Прежде всего, выдвигается *гипотеза* $\mathbf{H_0}$ о том, что распределение F удовлетворяет некоторому свойству. При этом, если есть возможность, желательно сразу конкретизировать альтернативное утверждение $\mathbf{H_1}$ относительно F. Например,

i) $\mathbf{H_0}: F = F_0$ – распределение F в точности совпадает с некоторым известным распределением F_0. Так, вполне уместно рассмотреть гипотезу о том, что случайные числа, выдаваемые функцией Random языка программирования Pascal, действительно имеют равномерное на отрезке [0;1] распределение;

ii) $\mathbf{H}_0: F \in \Psi$ – распределение F принадлежит некоторому известному семейству распределений Ψ. Весьма популярна на практике гипотеза о том, что наблюдаемая в эксперименте выборка имеет нормальное распределение с некоторыми неизвестными значениями среднего и дисперсии. Варианты альтернатив $\mathbf{H_1}$ как в этом, так и в предыдущем примере очень разнообразны и зависят от предпочтений исследователя. Наиболее естественная альтернатива здесь – семейство всех возможных распределений на числовой прямой (в предыдущем примере – на отрезке [0;1]);

iii) $\mathbf{H_0}: \theta \in \Theta_0$ – значение неизвестного параметра θ принадлежит некоторому подмножеству Θ_0. В качестве примера здесь можно привести задачу проверки гипотезы о том, что вероятность рождения мальчика больше ? , при альтернативе – меньше или равна ?. Другой пример, связанный с медициной, – исследование эффективности нового препарата при лечении пациентов с повышенным артериальным давлением в сравнении со стандартной методикой лечения. Здесь в качестве гипотезы лучше выдвинуть утверждение, что новый препарат идентичен старому, при альтернативе, что новый лучше старого.

♦ **Критерий.**

После выдвижения гипотезы строится критерий проверки этой гипотезы. *Критерий* – функция выборочных данных $\varphi(x^{(n)})$, принимающая значение

1, если гипотезу следует отвергнуть, и значение 0, если гипотезу следует принять. Вместо критерия часто строят *критическую область* A – область, при попадании в которую выборочных данных гипотеза отвергается.

Обычно критерий строится с помощью некоторой *тестовой статистики* $T(x^{(n)})$, и в этом случае гипотеза отвергается, если значение статистики больше (или меньше) *критической константы* $C_{\text{крит}}$. Величина константы выбирается в соответствии с требованиями на качество критерия.

♦ **Ошибки.**

Качество критерия (критической области) характеризуется двумя величинами – вероятностями ошибок первого и второго рода.

i) *Вероятность ошибки 1-ого рода* – вероятность отвергнуть гипотезу, если на самом деле она верна.

Вероятность ошибки 2-ого рода – вероятность принять гипотезу, если на самом деле верна альтернатива. Понятно, что если альтернатива не конкретизирована, то вычислить эту вероятность невозможно. Поэтому чаще всего контролируется только ошибка первого рода.

ii) Максимальная вероятность ошибки 1-го рода среди всех “гипотетических” распределений (распределений, удовлетворяющих требованиям проверяемой гипотезы) называется *размером* критерия.

iii) Критерий, размер которого не превосходит некоторого наперед заданного малого значения α, называется *критерием уровня* α, а заданная заранее константа α называется *уровнем значимости*.

Таким образом, для критерия уровня α маловероятно (не больше α) отвержение справедливой гипотезы:

$$\mathbf{P}_F\{\,\mathbf{H}_0 \text{ отвергается}\,\} \le \alpha \quad \text{для всех } F \in \mathbf{H}_0.$$

Поскольку на практике чаще всего контролируется только вероятность ошибки 1-го рода, то в качестве нулевой гипотезы следует выбирать утверждение, противоположное ожидаемому. Например, при исследовании нового лечебного препарата следует проверять гипотезу о его полной неэффективности. В этом случае критерий уровня α будет обеспечивать низкую вероятность принятия неэффективного препарата.

Чем меньше уровень значимости, тем выше надежность статистического вывода. Однако слишком малые значения α могут привести к слишком частому принятию неверной гипотезы, то есть к увеличению вероятности ошибки 2-ого рода. Чтобы контролировать обе эти ошибки, часто требуется проведение большого числа наблюдений, что не всегда допустимо.

Наиболее популярное на практике значение $\alpha = 0{,}05$ (- 5%-ый уровень значимости). Иногда рассматривают значения $\alpha = 0{,}10$ и $\alpha = 0{,}01$.

♦ **Тестовая статистика.**

Если критерий построен на основе статистики $T = T(X^{(n)})$ и критическая область имеет вид $T > C_{\text{крит}}$, то критическая константа выбирается из условия на вероятность ошибки 1-го рода

$$\mathbf{P}_F \{T(X^{(n)}) > C_{\text{крит}}\} \le \alpha, \qquad (*)$$

которое должно выполняться для всех распределений F, удовлетворяющих предположениям гипотезы $\mathbf{H_0}$. Если распределение статистики $T(X^{(n)})$ известно, то значение критической константы легко найти из таблиц этого распределения. Таким образом, для построения критерия проверки гипотезы необходимо

i) выбрать тестовую статистику T и вид критической области (например, $T > C_{\text{крит}}$);

ii) найти критическую константу $C_{\text{крит}}$ по таблице распределения T из условия (*) на вероятность ошибки 1-го рода;

iii) после проведения статистического эксперимента и вычисления значения $T = t_{\text{эксп}}$ принять (если $t_{\text{эксп}} \le C_{\text{крит}}$) или отвергнуть (если $t_{\text{эксп}} > C_{\text{крит}}$) проверяемую гипотезу.

Замечание. Не зная $C_{\text{крит}}$, нельзя проверить гипотезу, как бы не было соблазнительно значение T, полученное в эксперименте. Например, если из двух баскетболистов один попал в 25% бросков с игры, а второй – в 50%, то ещё нельзя сказать, что первый баскетболист хуже второго, так как

для подобного вывода надо знать (чтобы вычислить $C_{\text{крит}}$) количество бросков каждого из игроков.

♦ **Критический уровень значимости.**

Другой способ проверки гипотезы состоит в вычислении *критического уровня значимости*, который тесно связан с видом критерия и применительно к рассмотренному выше критерию равен

$$\alpha_{\text{крит}} = \sup_{F \in \mathbf{H_0}} \mathbf{P}_F\{T \geq t_{\text{эксп}}\}$$

(часто обозначается через p и называется p-*значением*).

Гипотеза $\mathbf{H_0}$ отвергается, если $\alpha_{\text{крит}} \leq \alpha$.

Другими словами, критический уровень значимости равен размеру критерия, при вычислении которого критическая константа $C_{\text{крит}}$ заменена полученным в эксперименте выборочным значением статистики $T = t_{\text{эксп}}$. Если вспомнить определение константы $C_{\text{крит}}$, то критический уровень значимости можно определить как

наименьший уровень значимости, при котором гипотеза отвергается.

В общем случае $\alpha_{\text{крит}}$ можно понимать как вероятность того, что тестовая статистика T примет значение “хуже”, чем полученное в эксперименте. Слово “хуже” означает, что такие значения статистики свидетельствуют больше в пользу альтернативы, чем гипотезы. Такое понимание поможет правильно сориентироваться в определении вида критической области ($T \geq t_{\text{эксп}}$, $T \leq t_{\text{эксп}}$ или $|T| \geq t_{\text{эксп}}$) при вычислении p-значения.

Основное преимущество этого способа (кроме очевидного преимущества, связанного с вычислением прямой, а не обратной функции распределения) состоит в том, что мы можем не делать жестких выводов типа “да, гипотеза верна” или “нет, гипотеза не верна”, а принять более гибкое решение. Здесь можно предложить следующую градацию высказываний о справедливости гипотезы:

$0.15 < \alpha_{\text{крит}}$ – хорошее согласие с гипотезой;

$0.10 \leq \alpha_{\text{крит}} \leq 0.15$ – нет оснований отвергать гипотезу;

$0.05 \le \alpha_{\text{крит}} \le 0.10$ – слабо значимое расхождение с гипотезой;

$0.01 \le \alpha_{\text{крит}} \le 0.05$ – значимое расхождение с гипотезой;

$\alpha_{\text{крит}} < 0.01$ – высоко значимое расхождение с гипотезой.

Следует предостеречь, что учет только ошибки 1-го рода может привести к весьма парадоксальным результатам. Можно предложить простой критерий, который будет иметь заданный уровень значимости сразу для всех гипотез. Для этого, например, достаточно 5 раз подбросить правильную монету и отвергать гипотезу, если все 5 раз монета упадет “гербом” вверх. Легко понять, что вероятность этого события равна $1/2^5$. Следовательно, размер такого (ну, очень смешного) критерия приблизительно равен 0.03, что, конечно, можно считать очень хорошим значением. Подчеркнем, что этот критерий может быть применен к проверке любой гипотезы, например, гипотезы о том, что на Луне есть жизнь. Чтобы избежать подобных нелепых ситуаций, необходимо либо применять критерии, у которых мала ошибка 2-го рода, вычисленная хотя бы для некоторых более или менее правдоподобных альтернатив, либо, в крайнем случае, использовать тестовые статистики, свидетельствующие в той или иной степени о близости данных к гипотезе. Например, если в эксперименте подсчитывается частота рождения мальчиков, то достаточно большое отклонение этой частоты от ? несомненно будет свидетельствовать против гипотезы о равенстве вероятностей рождения мальчиков и девочек.

§ 3. *Точечное оценивание*

Оценкой параметра θ, характеризующего распределение наблюдаемой в эксперименте выборки, называется статистика $\hat{\theta} = \hat{\theta}(x^{(n)})$, принимающая значения в пространстве возможных значений этого параметра. Среди свойств оценок выделяют обычно три основных –

- несмещенность;
- состоятельность;
- оптимальность.

♦ **Несмещенность.**

Оценка $\hat{\theta}$ называется *несмещенной*, если для всех значений неизвестного параметра θ

$$\mathbf{E}_\theta \hat{\theta} = \theta,$$

где, как обычно, индекс у знака математического ожидания означает, что вычисления производятся при истинном значении параметра, равном θ. Другими словами, несмещенность означает, что оценка, изменяясь от эксперимента к эксперименту, будет принимать значения, в среднем совпадающие с истинным значением оцениваемого параметра. Рассмотрим примеры.

i) Выборочное среднее $\overline{X}$ есть несмещенная оценка истинного среднего значения распределения $\mu = \mathbf{E}\, X$:

$$\mathbf{E}_\mu \overline{X} = \mu .$$

Вспомним по этому поводу известную русскую поговорку: “семь раз отмерь, вычисли среднее арифметическое и отрежь”.

ii) Относительная частота осуществления некоторого события A (количество ν появлений события в выборке, деленное на общее число выборочных данных n) есть несмещенная оценка вероятности события $p = \mathbf{P}\{A\}$:

$$\mathbf{E}_p \frac{\nu}{n} = p .$$

iii) Выборочная дисперсия S^2 – смещенная оценка истинной дисперсии $\sigma^2 = \mathbf{D}\, X = \mathbf{E}(X-\mu)^2$:

$$\mathbf{E}_{\sigma^2} S^2 = \frac{n-1}{n}\sigma^2 .$$

На практике чаще всего рассматривают так называемую исправленную на несмещенность выборочную дисперсию $\hat{S}^2 = \frac{n}{n-1} S^2$.

Для более сложных характеристик распределения в большинстве своем оценки получаются смещенными. В этом случае на применяемую оценку накладывают условие *асимптотической несмещенности*: $\lim\limits_{n\to\infty} \mathbf{E}_\theta \hat{\theta} = \theta$.

Этому требованию удовлетворяют уже почти все используемые на практике оценки.

♦ **Состоятельность.**

Оценка $\hat{\theta}$ называется *состоятельной* оценкой параметра θ, если при истинном значении параметра, равном θ, и при объеме выборки $n \to \infty$

$$\hat{\theta} \to \theta .$$

Другими словами, состоятельная оценка с ростом объема выборки приближается к истинному значению оцениваемого параметра. В случае, если имеет место сходимость по вероятности, говорят о слабой состоятельности, при сходимости почти наверное говорят о сильной состоятельности.

Большинство оценок вычисляется либо непосредственно как сумма некоторых функций от выборочных наблюдений, либо как непрерывная функция от таких сумм. Для исследования состоятельности подобных оценок полезна теорема, известная в теории вероятностей как Закон Больших Чисел, утверждающая, что среднее арифметическое с ростом числа слагаемых стремится (по вероятности или почти наверное) к теоретическому среднему своих слагаемых. Все вышерассмотренные оценки будут состоятельными именно в силу закона больших чисел. В качестве дополнительного примера можно еще привести выборочное стандартное отклонение $S = \sqrt{S^2}$. Эта оценка, очевидно, будет состоятельной для истинного стандартного отклонения σ.

♦ **Оптимальность.**

Точность оценок характеризуется величиной риска. Наиболее исследовано сравнение оценок на основе *среднеквадратического риска*:

$$R(\hat{\theta} \mid \theta) = \mathbf{E}_\theta (\hat{\theta} - \theta)^2 .$$

Для несмещенных оценок риск есть не что иное, как дисперсия оценки.

Понятно, что предпочтительнее выбирать оценки, у которых риск принимает наименьшее значение среди всех допустимых оценок. К сожалению, это трудно достижимо, а точнее – никогда не достижимо. В

качестве доказательства этого положения рассмотрим оценку $\hat{\theta} \equiv 7.5$, которая, конечно, имеет плохой риск при $\theta \neq 7.5$, однако в точке $\theta = 7.5$ её риск равен нулю, что невозможно улучшить. В связи с этим задачу построения оценки с наилучшим риском рассматривают обычно в классе оценок, удовлетворяющих некоторому дополнительному свойству, например, свойству несмещенности. Доказано, что для большинства практически полезных вероятностных моделей все рассмотренные выше несмещенные оценки будут иметь минимальную дисперсию.

Интересно отметить здесь, что с точки зрения среднеквадратического риска качество смещенной оценки дисперсии S^2 выше качества её несмещенного варианта $\hat{S}^2$. Так что, как поется в известной песне, – "думайте сами, решайте сами иметь, несмещенность или не иметь".

§ 4. *Доверительное оценивание.*

Точечная оценка несет определенную информацию о величине неизвестного параметра. Естественно, здесь возникает вопрос о точности этой оценки. Ответ на этот вопрос дают так называемые интервальные оценки или доверительные множества.

Пусть $x^{(n)} = (x_1, \ldots, x_n)$ – случайная выборка, распределение которой $F_\theta(x)$ зависит от некоторого неизвестного параметра θ. Интервал $(\underline{\theta}; \overline{\theta})$ с границами $(\underline{\theta}(x^{(n)}); \overline{\theta}\ (x^{(n)}))$, зависящими от выборочных данных, называется $(1-\alpha)$-*доверительным интервалом* для параметра θ, если

$$\mathbf{P}_\theta\{\underline{\theta} < \theta < \overline{\theta}\} \geq 1-\alpha. \qquad (1)$$

Величина $Q = (1\text{-}\alpha)\cdot 100\%$ называется *надежностью* интервала и выбирается обычно в пределах от 90% до 99% (стандартное значение – 95%).

Статистика $\overline{\theta}$ называется *верхней* $(1-\alpha)$-*доверительной границей* для параметра θ, если

$$\mathbf{P}_\theta\{\theta < \overline{\theta}\} \geq 1-\alpha.$$

Статистика $\overline{\theta}$ называется *нижней* $(1-\alpha)$-*доверительной границей* для параметра θ, если

$$\mathbf{P}_\theta\{\underline{\theta} < \theta\} \geq 1-\alpha\,.$$

Если указанные в этих определениях неравенства выполняются в пределе при увеличении объема выборки до бесконечности, то такие доверительные интервалы называются *асимптотическими*.

Более подробно вопрос построения доверительных множеств рассматривается в главе, посвященной интервальным оценкам.

§ 5. *Вероятностные модели*

Как видно из вышеизложенного, качество любой оценки и любого критерия можно выяснить, только если известно точное или асимптотическое распределение соответствующей статистики. Здесь мы опишем наиболее часто встречающиеся на практике модели распределений.

♦ **Распределение.**

Функцией распределения случайной величины X называется функция

$$F(x) = \mathbf{P}\{X < x\},$$

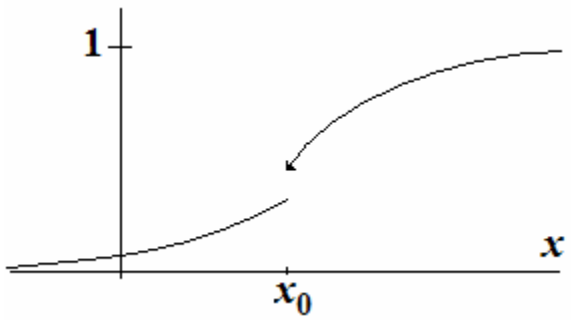
то есть вероятность попадания левее фиксированного значения x.

Функция распределения изменяется от 0 (при $x \to -\infty$) до 1 (при $x \to +\infty$), нигде не убывает и непрерывна слева.

Скачок функции распределения возможен лишь в точке x_0, вероятность попадания в которую с.в. X не равна нулю:

$$\mathbf{P}\{X = x_0\} = p_0 > 0\,.$$

Величина скачка равна p_0.



Функция $\overline{F}(x) = 1 - F(x)$, то есть вероятность попадания с.в. X правее значения x, называется *функцией надежности*. Это определение связано с тем, что, если с.в. описывает время службы некоторого прибора,

то функция $\overline{F}(x)$ равна вероятности того, что этот прибор прослужит дольше заданного времени x.

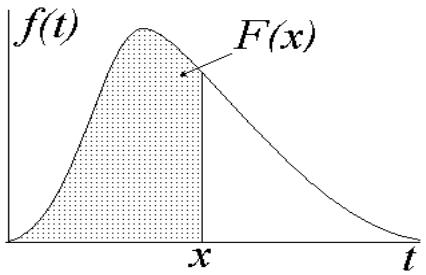
Если функция распределения представима в виде интеграла $F(x) = \int_{-\infty}^{x} f(t)\,dt$, то распределение называется абсолютно-непрерывным, а функция $f(x)$ *функцией плотности*. Для практически полезных распределений плотность равна производной функции распределения: $f(x) = F'(x)$.

Функция плотности $f(x)$ обладает следующими свойствами:

(i) $f(x) \geq 0$ и полный интеграл $\int_{-\infty}^{\infty} f(x)\,dx = 1$;

(ii) если функция f четна, то распределение симметрично около нуля: $F(-x) = 1 - F(x)$, иными словами, график функции распределения симметричен около точки (0,?).

(iii) значение функции распределения $F(x)$ равно площади под графиком функции плотности в интервале от $-\infty$ до x:



Свойство *(ii)* часто используют при построении таблиц распределения, ограничиваясь лишь положительными значениями аргумента.

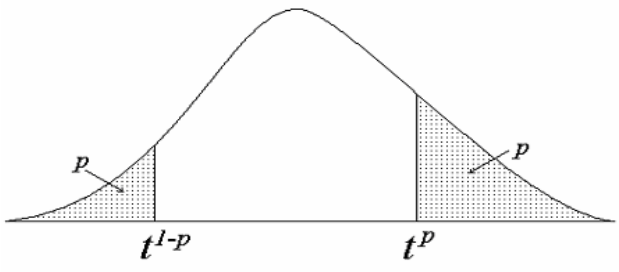
♦ **Квантили – процентные точки.**

Решение уравнения $F(t) = p$ называют p-*квантилью* распределения F, а уравнения $1 - F(t) = p$ *верхней* p-*квантилью*.

Если обозначить верхнюю p-квантиль через t^p, то, очевидно, p-квантиль будет равна t^{1-p}. Если функция распределения имеет обратную, то

$$t^p = F^{-1}(1-p) = \overline{F}^{-1}(p).$$

Графически квантили распределения можно представить следующим образом (см. свойство плотности *(iii)*):



Для симметричных распределений $t^{1-p} = -t^p$. Поэтому для таких распределений таблицы составляются лишь для значений $p < 1/2$.

Верхние квантили для вероятностей, выраженных в процентах ($Q = p \cdot 100\%$), часто называют *процентными точками* распределения.

♦ **Нормальное распределение.**

Случайная величина X имеет *нормальное распределение* с параметрами (μ, σ^2) (коротко $X \sim N(\mu, \sigma^2)$), если её функция распределения равна

$$F(x) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) dt .$$

Плотность нормального распределения равна

$$f(x) = F'(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

Параметр μ равен среднему значению с.в. X: $\mu = \mathbf{E}\, X$.

Параметр σ^2 равен дисперсии с.в. X: $\sigma^2 = \mathbf{D}\, X = \mathbf{E}(X - \mu)^2$.

Широкую распространенность нормального закона в природе можно объяснить, исходя из Центральной Предельной Теоремы. Представим себе, что мы многократно измеряем некоторую характеристику изучаемого объекта посредством очень точного прибора. Ясно, что результат будет случайно изменяться от измерения к измерению. Случайность этого изменения обусловлена суммарным влиянием большого количества факторов. Если каждый из этих факторов не оказывает решающего влияния на окончательный результат измерения, то можно предположить, что показания прибора – суть реализации нормальной случайной величины.

Нормальный закон с параметрами $\mu = 0$ и $\sigma^2 = 1$ (то есть $X \sim N(0,1)$) называется *стандартным*, его функция распределения обозначается $\Phi(x)$ ("фи" или, по-старорусски, "ферт"), а функция плотности $\varphi(x)$. Общая функция распределения нормального закона может быть записана через стандартную: $F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$. Аналогично записывается общая функция плотности нормального закона: $f(x) = \frac{1}{\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right)$.

Практически любой справочник по математической статистике содержит таблицы функции $\Phi(x)$ и её квантилей. Так как стандартное нормальное распределение симметрично, то эти таблицы составляют для значений $x \geq 0$ и $p < 1/2$. Приведем фрагмент таблицы из сборника [1].

Слева в таблице представлено входное значение x с точностью до второго знака после запятой. Третий знак указан в самой верхней строке таблицы. С целью представления на одном листе по возможности большей информации, таблица разбита на блоки (отделенные чертой), в которых числа имеют несколько одинаковых первых цифр. Эти совпадающие части приведены только для одного значения (в столбце под верхней первой ячейкой с цифрой 0).

Таблица 1.1. Функция нормального распределения $\Phi(x)$

x	0	1	2	3	4	…	8	9
				…				
2,05	0,97 9818	9867	9915	9964	0012	…	0205	0253
06	0,98 0301	0349	0396	0444	0491	…	0680	0727
07	0774	0821	0867	0914	0960	…	1145	1191
				…				

Так, например,

$$\Phi(2{,}052) = 0{,}97\,9915, \quad \Phi(2{,}058) = 0{,}98\,0205, \quad \Phi(2{,}074) = 0{,}98\,0960.$$

Для нахождения квантилей можно использовать таблицу исходной функции распределения, отыскивая значение вероятности внутри таблицы и находя соответствующее ему входное значение. Например, при $p = 0{,}02$ верхняя p-квантиль (то есть решение уравнения $\Phi(x) = 1 - p = 0{,}98$) будет

находиться где-то между 2,053 и 2,054, так как $\Phi(2{,}053) < 0{,}98 < \Phi(2{,}054)$. Простая линейная аппроксимация дает $t^{0{,}02} \approx 2{,}0537$.

В этом же сборнике [1] имеется таблица значений обратной функции нормального распределения, иными словами – таблица *p*-квантилей (не верхних).

Таблица 1.3. Функция, обратная функции нормального распределения

p	0	1	2	3	4	…	8	9
				…				
977	1,9 9539	9723	9908	0093	0279	…	1030	1219
978	2,0 1409	1600	1792	1984	2177	…	2957	3154
979	3352	3551	3750	3950	4151	…	4964	5169
0,980	5375	5582	5790	6000	6208	…	7056	7270
				…				

Таким образом, $t^{0{,}02} = \Phi^{-1}(0{,}98) = 2{,}05375$, что весьма близко к полученному выше приближенному значению.

♦ **Хи-квадрат распределение.**

Если с.в. можно представить в виде суммы квадратов m независимых стандартных нормальных с.в. –

$$\xi_1^2 + \xi_2^2 + \ldots + \xi_m^2,$$

то эта с.в. будет иметь *распределение хи-квадрат с* m *степенями свободы* (число степеней свободы совпадает с числом слагаемых). Её принято обозначать символом χ_m^2. Для особо любознательных скажем, что хи-квадрат распределение совпадает с гамма-распределением с параметрами $(m/2, 1/2)$. Функция распределения χ_m^2 выражается формулой

$$K_m(x) = \mathbf{P}\{\chi_m^2 < x\} = \int_0^x \frac{1}{2^{m/2}\Gamma(m/2)} y^{m/2-1} \exp(-\tfrac{y}{2})\, dy, \quad x > 0,$$

где $\Gamma(p)$ – гамма-функция Эйлера; если $x \le 0$, то $K_m(x) = 0$.

Сборник таблиц [1] содержит значения так называемого *интеграла вероятностей хи-квадрат* – в нашей терминологии это просто функция надежности $\overline{K}_m(x) = 1 - K_m(x)$. Таблица имеет два входа – по числу

степеней свободы (верхняя строка) и по аргументу функции (левый столбец).

Таблица 2.1а. Интеграл вероятностей χ^2

x	m=16		…	m=20	
	P	-Δ		P	-Δ
…	…			…	
15,0	0,52464	3627	…	0,77641	2929
5	48837	3541	…	74712	3050
…	…			…	

Здесь, кроме значения функции распределения (столбец P), приведены также первые разности этой функции (столбец -Δ), точнее, только 5 значащих цифр после запятой без первых нулей. Таким образом, $\overline{K}_{16}(15{,}5) - \overline{K}_{16}(15{,}0) = -0{,}03627$ (после запятой поставлен один ноль, чтобы получилось пять цифр). Если x_0 и x_1 – два рядом стоящие значения аргумента, то для нахождения значения функции в промежуточной точке $x \in [x_0; x_1]$ можно применить аппроксимацию $\overline{K}_m(x) \approx \overline{K}_m(x_0) - \Delta \cdot \frac{x - x_0}{x_1 - x_0}$. В приведенном нами фрагменте $x_1 - x_0 = 0{,}5$. Поэтому $\overline{K}_{16}(15{,}2) \approx \overline{K}_{16}(15{,}0) - 0{,}03627 \cdot 2 \cdot 0{,}2 = 0{,}510132$.

Значения верхних p-квантилей $t^p = t^p(m) = \overline{K}_m^{-1}(p)$ распределения хи-квадрат содержатся в следующей таблице на стр.166 сборника [1].

Таблица 2.2а. Процентные точки распределения χ^2

m \ Q	…	97,5%	95%	…	5%	2,5%	…
…				…			
19	…	8,907	10,117	…	30,144	32,852	…
20	…	9,591	10,851	…	31,410	34,170	…
…				…			

Вход в таблицу осуществляется по числу степеней свободы (m в левом столбце) и по вероятности, выраженной в процентах (Q в верхней строке). Таким образом, $t^{0{,}05}(19) = 30{,}144$, $t^{0{,}025}(20) = 34{,}170$.

- **Распределение Стьюдента.**

Если с.в. T_k можно представить в виде отношения $T_k = \dfrac{\xi}{\sqrt{\chi_k^2}}\sqrt{k}$,

где ξ – стандартная нормальная (0,1) с.в.,

χ_k^2 – хи-квадрат с.в. с k степенями свободы, не зависящая от ξ,

то T_k будет иметь *распределение Стьюдента с k степенями свободы*. Функция распределения Стьюдента имеет вид

$$S_k(t) = \mathbf{P}\{T_k < t\} = \int_{-\infty}^{t} \frac{\Gamma((n+1)/2)}{\sqrt{k\pi}\,\Gamma(n/2)}\left(1+\frac{u^2}{k}\right)^{-\frac{(k+1)}{2}} du,$$

где $\Gamma(p)$ – гамма-функция Эйлера. Так как функция плотности этого распределения (подынтегральная функция) четна, то распределение Стьюдента симметрично: $S_k(-t) = 1 - S_k(t)$.

Легко видеть, что, так как с.в. χ_k^2 есть сумма квадратов стандартных нормальных с.в., то отношение χ_k^2 / k стремится по вероятности к 1. Поэтому при большом значении степени свободы k распределение Стьюдента может быть аппроксимировано нормальным (0,1) распределением (как у числителя ξ).

Таблицы распределения Стьюдента также имеются в любом справочнике по математической статистике. Приведем здесь фрагмент соответствующей таблицы из сборника [1].

Таблица 3.1а. Функция распределения Стьюдента

t \ k	11	12	…	19	20
…			…		
2,0	0,9646	0,9657	…	0,9700	0,9704
1	9702	9712	…	9753	9757
…			…		

Эта таблица имеет два входа – число степеней свободы k (верхняя строка) и аргумент функции t (левый столбец). Из этой таблицы находим, что $S_{12}(2) = 0{,}9657$, $S_{19}(2{,}1) = 0{,}9753$. При степенях свободы больше 20 можно воспользоваться нормальным приближением: $S_k(t) \approx \Phi(t), k > 20$.

Следующая таблица указанного сборника [1] содержит значения верхних p-квантилей $t^p = t^p(k) = \overline{S}_k^{-1}(p)$. Эта таблица также имеет два входа – число степеней свободы (левый столбец) и вероятность в процентах $Q = p \cdot 100\%$ (верхняя строка). Для наглядности целые части вместе с запятой приведены только для верхних чисел в блоке из пяти чисел.

Таблица 3.2. Процентные точки распределения Стьюдента

k \ *Q*	…	10%	5%	2,5%	…	0,05%
…			…			
19	…	1,3277	1,7291	2,0930	…	3,8834
20	…	3253	7247	0860	…	8495
…			…			

Таким образом, $t^{0,1}(16) = 1{,}3368, \quad t^{0,025}(17) = 2{,}1098$.

♦ **Показательное (экспоненциальное) распределение.**

Если про некоторый объект (прибор) можно утверждать, что вероятность выхода его из строя за время t не зависит от того, сколько этот объект уже прослужил (эффект отсутствия последействия), то общее время службы такого объекта T будет подчинено *показательному закону* с функцией распределения

$$F(t) = \mathbf{P}\{T < t\} = 1 - e^{-t/\lambda},\ t \geq 0.$$

Функция плотности показательного закона равна

$$f(t) = \tfrac{1}{\lambda} e^{-t/\lambda},\ t \geq 0.$$

Коэффициент $\lambda\ (> 0)$ называется интенсивностью отказа и несет информацию о количестве вышедших из строя объектов в единицу времени среди всех объектов, прослуживших данное время t. Показательное распределение имеет постоянную (не зависящую от срока службы t) интенсивность отказа.

Для показательного распределения характерно совпадение среднего значения и стандартного отклонения:

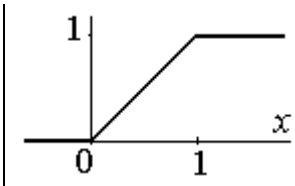
$$\lambda = \mathbf{E}\,X = \sqrt{\mathbf{D}\,X}\,.$$

Это свойство часто используют для проверки согласия выборочных данных с экспоненциальной моделью распределения.

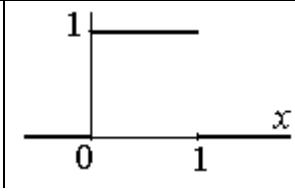
♦ **Равномерное распределение на [0; 1].**

С.в. X, для которой вероятность попадания в любой интервал пропорциональна длине этого интервала, имеет равномерное распределение. Чаще всего рассматривают равномерную на [0; 1] с.в. В этом случае функция распределения и функция плотности X равны

$$F(x) = \begin{cases} 0, & \text{åñëè } x \le 0, \\ x, & \text{åñëè } 0 \le x \le 1, \\ 1, & \text{åñëè } x \ge 1, \end{cases}$$



$$f(x) = \begin{cases} 0, & \text{åñëè } x < 0 \text{ èëè } x > 1, \\ 1, & \text{åñëè } 0 \le x \le 1. \end{cases}$$



Равномерное распределение есть обобщение на непрерывный случай классической схемы выбора из конечной популяции. Так, например, датчик случайных чисел любого языка программирования в действительности выбирает с равной вероятностью целые числа k от 0 до N (скажем, $N = 10^9$). “Непрерывные” равномерные числа X из отрезка [0; 1] получаются путем простого преобразования $X = k / N$.

♦ **Биномиальное распределение.**

Рассмотрим n независимых экспериментов, в каждом из которых наблюдается некоторое фиксированное событие A (например, поражение цели при стрельбе). Если вероятность этого события p не изменяется от эксперимента к эксперименту, то сл.в. X, равная числу успешных реализаций события A, имеет *биномиальное распределение* с параметрами (n, p):

$$\mathbf{P}\{X < x\} := Bin(x \mid n, p) = \sum_{k=0}^{x-1} C_n^k p^k (1-p)^{n-k}, \quad x = 0,1,\ldots, n+1.$$

При больших значениях n и не очень маленьких значениях $p(1-p)$ биномиальное распределение может быть аппроксимировано нормальным:

$$Bin(x \mid n, p) \approx \Phi\left(\frac{x-np}{\sqrt{np(1-p)}}\right).$$

Другими словами, можно сказать, что биномиальная с.в. X имеет приближенно нормальное распределение с параметрами $(np, np(1-p))$.

При больших значениях n и малых значениях p (или $1-p$) биномиальное распределение аппроксимируется пуассоновским распределением:

$$Bin(x \mid n, p) = e^{-\lambda} \sum_{k=0}^{x-1} \frac{\lambda^k}{k!}, \quad x = 0,1,\ldots, \text{ ãäå } \lambda = np.$$

Для биномиальной с параметрами (n, p) с.в. X

среднее значение равно np,

дисперсия равна $np(1-p)$.

Контрольные вопросы.

1. Что такое функция распределения (функция плотности)?
   Ответ: см. стр. 14.
2. Что такое функция надежности?
   Ответ: см. стр. 14.
3. Какая случайная величина имеет нормальное распределение (показательное, равномерное, биномиальное, хи-квадрат, Стьюдента)?
   Ответ: см. стр. 16-22.
4. Запишите формулу для функции распределения нормального закона (экспоненциального, равномерного, биномиального).
   Ответ: см. стр. 16-22.
5. Какой смысл несут параметры нормального распределения (экспоненциального, биномиального, хи-квадрат, Стьюдента)?
   Ответ: см. стр. 16-22.
6. Чему равны среднее значение и дисперсия экспоненциального распределения (нормального, биномиального)?
   Ответ: см. стр. 16-22.
7. Что такое квантиль (верхняя квантиль) распределения?
   Ответ: см. стр. 15.
8. Как связаны функция распределения и её верхняя квантиль?
   Ответ: см. стр. 15.
9. Найдите по таблице значение верхней 7%-й квантили для распределения хи-квадрат при 15 степенях свободы (для нормального распределения, для распределения Стьюдента).
   Ответ: см. стр. 17-22.
10. Что такое выборка?
    Ответ: см. стр. 5.
11. Что такое оценка?
    Ответ: см. стр. 10.
12. Дайте определение состоятельности оценки и проинтерпретируйте смысл этого определения.
    Ответ: см. стр. 12

13. Можно ли сказать, что состоятельная оценка лучше не состоятельной оценки?
14. Дайте определение несмещенности оценки и проинтерпретируйте смысл этого определения.
    Ответ: см. стр. 10.
15. Можно ли сказать, что несмещенная оценка лучше смещенной оценки?
16. Как следует выбирать нулевую гипотезу?
    Ответ: см. стр. 7.
17. Как определяется вероятность ошибки 1-го рода? Что такое размер критерия?
    Ответ: см. стр. 7.
18. Что такое уровень значимости?
    Ответ: см. стр. 7.
19. Какой уровень значимости лучше выбрать – 5% , 10% или 1%?
    Ответ: см. стр. 7-8.
20. Как часто мы будем ошибаться, если будем применять критерий уровня $\alpha = 0{,}03$.
    Ответ: см. стр. 7.
21. Как построить критерий заданного уровня, основываясь на значениях некоторой статистики T?
    Ответ: см. стр. 8.
22. Можно ли признать новый метод лечения лучше старого, если при клинических испытаниях результативность нового метода составила 85%, а старого – 70%? Что ещё нужно знать, что бы правильно ответить на этот вопрос?
    Ответ: см. стр. 8.
23. Что такое критический уровень значимости? Чем он отличается от уровня значимости?
    Ответ: см. стр. 7, 9.
24. Как проверить гипотезу, основываясь на значении критического уровня значимости?
    Ответ: см. стр. 9.

Глава II. Первичный статистический анализ

В этой главе приводятся описания первых трех заданий. Рассматриваемые здесь процедуры предваряют, как правило, любую статистическую обработку данных.

Задание 1.

Выборочные характеристики.

Постановка задачи.

Вычислить основные статистические характеристики выборочных данных:

- Среднее арифметическое.
- Дисперсию.
- Стандартное отклонение.
- Коэффициент асимметрии.
- Коэффициент эксцесса.

Теоретические основы.

Статистический анализ выборочных данных начинают обычно с вычисления выборочных моментов.

- *Среднее*

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

(читается “икс с чертой”) – несмещенная и состоятельная оценка истинного среднего значения μ. Величина μ характеризует расположение наблюдаемой с.в. X. Очень часто сравнение различных совокупностей производят как раз по среднему. Однако, надо иметь в виду, что это имеет смысл только в случае, если, во-первых, плотность распределения X имеет

один локальных максимум и, во-вторых, дисперсии наблюдаемых с.в. (см. ниже) во всех группах "почти" одинаковы.

Дисперсия

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \bar{x}^2$$

– состоятельная и "почти несмещенная" оценка истинной дисперсии σ^2. Служит мерой разброса с.в. около её среднего μ. По известному правилу трех сигм с вероятностью, большей 90%, следует ожидать значение с.в. в пределах $\mu \pm 3\sigma$. Интересно, что нормальная с.в. с вероятностью, большей 95%, принимает значения в интервале $\mu \pm 2\sigma$.

Дисперсия измерительного прибора характеризует его точность.

Наряду с дисперсией, всегда вычисляется

стандартное отклонение $s = \sqrt{s^2}$.

Забегая вперед, скажем, что точность доверительного интервала для истинного значения среднего μ прямо пропорциональна s. Грубо говоря, истинное μ лежит где-то в пределах $\pm\frac{2}{\sqrt{n}}s$ от выборочного среднего $\bar{x}$.

Коэффициент асимметрии

$$g_1 = \frac{1}{ns^3}\sum_{i=1}^{n}(x_i - \bar{x})^3 = \frac{1}{s^3}\left(\frac{1}{n}\sum_{i=1}^{n}x_i^3 - 3\bar{x}\,\frac{1}{n}\sum_{i=1}^{n}x_i^2 + 2\bar{x}^3\right)$$

– состоятельная, но смещенная оценка истинного коэффициента асимметрии $\gamma_1 = E\left(\frac{X-\mu}{\sigma}\right)^3$. Несет информацию о симметричности расположения данных относительно центра $\bar{x}$. При больших положительных значениях γ_1 распределение с.в. будет иметь более "тяжелый" правый "хвост". Если график плотности такого распределения "насадить" на вертикальный штырь в точке достижения максимума, то график "упадет" вправо.

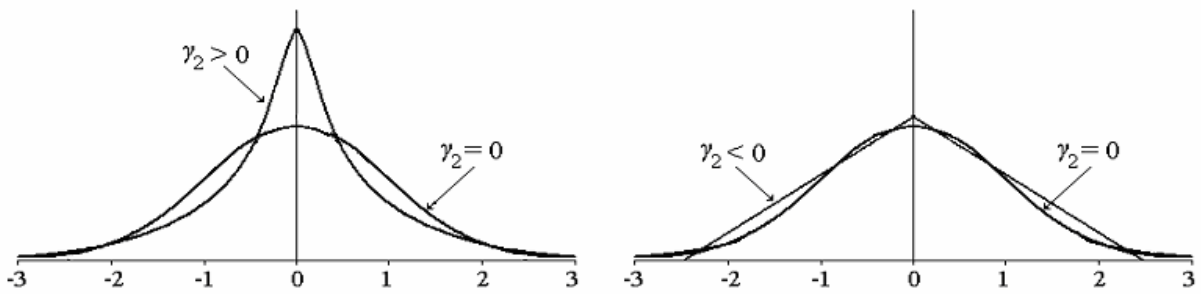
Выборочный коэффициент g_1 можно использовать для проверки согласия данных выборочного обследования с нормальной моделью распределения. Большие абсолютные значения g_1 будут свидетельствовать против гипотезы нормальности распределения, каковое, как известно, симметрично. При объеме выборки n, близком к 100, и

уровне значимости $\alpha = 0{,}10$ гипотезу нормальности следует отвергать, если выборочный коэффициент асимметрии g_1 не попадает в интервал (-0,389; 0,389) (см. сборник таблиц [1], стр.258). Границы этого интервала построены так, чтобы вероятность ошибки 1-ого рода, то есть вероятность отвержения гипотезы нормальности в случае, когда она верна, не превосходила заданного уровня α.

Коэффициент эксцесса

$$g_2 = \frac{1}{ns^4}\sum_{i=1}^{n}(x_i - \bar{x})^4 - 3 = \frac{1}{s^4}\left(\frac{1}{n}\sum_{i=1}^{n}x_i^4 - 4\bar{x}\,\frac{1}{n}\sum_{i=1}^{n}x_i^3 + 6\bar{x}^2\,\frac{1}{n}\sum_{i=1}^{n}x_i^2 - 3\bar{x}^4\right) - 3$$

– состоятельная, но смещенная оценка истинного коэффициента эксцесса $\gamma_2 = E\left(\frac{X-\mu}{\sigma}\right)^4 - 3$. Служит мерой сосредоточенности данных около среднего. У нормального распределения $\gamma_2 = 0$ (для этого и вычитается тройка). При положительных значениях γ_2 плотность распределения будет иметь более острый, чем у нормального распределения, пик около среднего. Соответственно, при отрицательных γ_2 пик будет более плоским. Иногда, по аналогии с видом нормальной функции плотности, коэффициент γ_2 называют коэффициентом “колоколообразности”.



На этом рисунке приведены графики трех плотностей:

1) нормального распределения – $\gamma_2 = 0$,

2) треугольного распределения – $\gamma_2 < 0$, и

3) двухстороннего экспоненциального распределения – $\gamma_2 > 0$.

Выборочный коэффициент g_2 также можно использовать для проверки гипотезы нормальности. При объеме выборки n, близком к 100, и уровне значимости $\alpha = 0{,}10$ гипотезу нормальности следует отвергать,

если коэффициент эксцесса g_2 не попадает в интервал (-0,65; 0,77) (см. сборник таблиц [1], стр.259).

Задание 2.

Гистограмма выборки.

Постановка задачи.

Построить график гистограммы выборки с подогнанной ожидаемой функцией плотности.

Теоретические основы.

- *Гистограмма* – ступенчатая кривая, высота ступенек которой пропорциональна количеству выборочных данных, попавших в заданные интервалы числовой прямой. Гистограмма используется для геометрического представления данных. В некотором смысле её можно считать оценкой функции плотности. Вероятность попадания в выбранный интервал может быть оценена относительной частотой попадания в этот интервал. Поэтому (и в силу теоремы о среднем значении) относительная частота, деленная на длину интервала, является оценкой функции плотности в некоторой средней точке этого интервала:

$$\frac{\nu}{n\Delta} \approx \frac{1}{\Delta}\int_{a}^{a+\Delta} f(x)\,dx = f(x_{\text{средн}}), \quad x_{\text{средн}} \in (a; a+\Delta).$$

В дальнейшем на основе гистограммы будет построен критерий проверки гипотезы согласия с гипотетическим распределением.

Для построения гистограммы необходимо

I) разбить область значений выборки на заданное число k интервалов (считая оба бесконечных крайних интервала);

II) для каждого j=1,…,k подсчитать количество ν_j выборочных данных, попавших в j -ый интервал;

III) построить график ступенчатой кривой, у которой высота ступеньки над j -ым интервалом пропорциональна ν_j;

IV) наложить на график гистограммы график функции плотности предполагаемого распределения (например, нормального).

Разберем по порядку каждый из этих пунктов.

I) Чаще всего все внутренние конечные интервалы выбираются одинаковой длины. Поэтому для построения разбиения достаточно сначала выбрать правую границу 1-го интервала a_1 и левую границу a_{k-1} последнего k-го интервала. Остальные границы вычисляются по формуле

$$a_j = a_1 + \Delta * j, \quad j = 2,...,k-1,$$

с шагом $\Delta = (a_{k-1} - a_1)/(k-2)$. Левая граница 1-го интервала равна $a_0 = -\infty$, правая граница последнего интервала $a_k = +\infty$. По поводу выбора количества интервалов k и первой границы a_1 существует множество различных мнений. Здесь необходимо учитывать как качество визуального представления, так и дальнейшее использование этой гистограммы для проверки гипотез и оценки вероятностей попадания наблюдаемой случайной величины в различные области. Если гистограмма используется только для визуального сравнения с функцией плотности, то, как вариант, можно взять число интервалов k равным приблизительно десятой части выборки, первую границу $a_1 = x_{\min} + \Delta/2$, а последнюю границу $a_{k-1} = x_{\max} - \Delta/2$, где $x_{\min}$ – минимальное, $x_{\max}$ – максимальное значения выборки, длина внутренних интервалов $\Delta = (x_{\max} - x_{\min})/(k-1)$. Кроме того, удобнее выбирать границы так, чтобы они не совпадали с данными. Для этого можно, например, границы задавать с большей точностью, чем точность выборки. В целях упрощения, количество интервалов, первая граница и длина интервалов будут заданы каждому из студентов отдельно (или в файле персональных данных, или преподавателем).

II) Подсчет количества попаданий в каждый интервал можно осуществить без использования компьютера. Для этого необходимо на листе бумаги начертить схему расположения интервалов (можно без соблюдения масштаба) и последовательно просмотреть все данные. При попадании очередного числа в j -ый интервал нужно поставить над этим интервалом точку. Количество точек над каждым из интервалов по окончании просмотра и будет равно искомой частоте. Для автоматизации

подсчета частот в Excel можно воспользоваться встроенной функцией Частота.

III) При построении гистограммы высоту ступеньки необходимо выбирать равной $\nu_j/(n\Delta)$, что даст в результате кривую, площадь под которой равна единице – аналог свойства функции плотности. Поскольку всё равно приходится выбирать некоторый масштаб представления графика, то в случае совпадения длин всех внутренних интервалов можно положить высоту ступеньки равной ν_j.

IV) Как уже было сказано, гистограмма представляет собой некую оценку функции плотности. Поэтому естественно попытаться сопоставить график гистограммы с графиком ожидаемой плотности $f(x)$. В идеале, если данные получены из распределения с плотностью $f(x)$, то график функции $f(x)$ должен пересечь каждую ступеньку графика гистограммы.

Чаще всего ожидается нормальное распределение. Из других популярных распределений мы выделим показательное распределение времени службы и равномерное на отрезке [0;1]. Так как обычно предполагаемое распределение зависит от некоторых неизвестных параметров, то при построении графика плотности необходимо эти параметры оценить. Для нормального закона неизвестное среднее μ оценивается выборочным средним $\overline{x}$, а дисперсия σ^2 – выборочной дисперсией s^2. Для показательного закона интенсивность отказа λ равна среднему значению, поэтому λ также может быть оценено посредством $\overline{x}$.

При одновременном прорисовывании графика гистограммы и функции плотности следует помнить о выбранном масштабе представления. Другими словами, если при вычислении высоты ступеньки мы забыли поделить на $n\Delta$, то придется на константу $n\Delta$ умножить каждое значение функции плотности: $n\Delta * f(x)$. Чтобы построить по возможности более точный график плотности, значения функции плотности необходимо было бы вычислить в большом числе точек. Однако возможности пакета Excel позволяют построить график плотности с достаточной степенью точности, если эти вычисления произвести только в

средних точках выбранных интервалов, после чего применить опцию сглаживания.

Задание 3.

Эмпирическая функция распределения.

Постановка задачи.

Построить график эмпирической функции распределения с подогнанной ожидаемой функцией распределения.

Теоретические основы.

- *Эмпирическая функция распределения*:

$$F_n(x) = \frac{\text{число выборочных данных } x_i, \text{ для которые } x_i < x}{n}.$$

Она служит оценкой истинной функции распределения и представляет собой возрастающую от 0 до 1 ступенчатую функцию. Изменения $F_n(x)$ происходят скачком в точках x, совпадающих с каким-либо выборочным значением x_i. Высота этого скачка равна числу выборочных данных, равных x_i, поделенному на общий объем выборки n. Внутри любого интервала значений x, не содержащего выборочных данных, функция $F_n(x)$ остается неизменной.

В отличие от гистограммы, ЭФР является достаточной статистикой, то есть сохраняет всю полноту информации выборки. Кроме того, при увеличении объема выборки она сходится к истинной функции распределения $F(x)$:

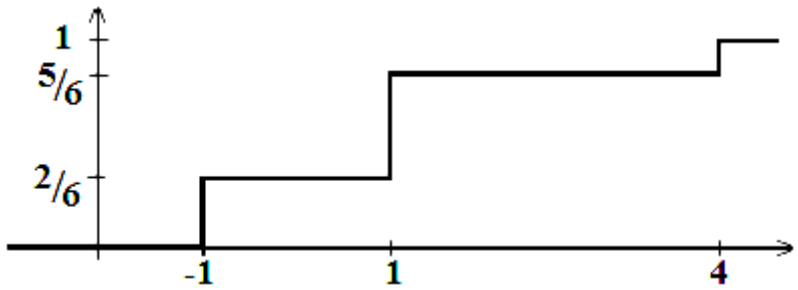
$$D = \sup_x |F_n(x) - F(x)| \underset{P}{\to} 0, \; n \to \infty. \qquad (*)$$

Другими словами, она является состоятельной оценкой $F(x)$. Легко показать, что ЭФР также и несмещенная оценка $F(x)$.

Для построения ЭФР необходимо сначала построить так называемый *вариационный ряд* – ряд упорядоченных выборочных данных $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$. Обратим внимание здесь на различие в написании индексов в исходной выборке x_i и в вариационном ряду $x_{(j)}$. В первом случае индекс указывает на номер в порядке поступления выборочного

значения, а во втором случае – на его ранг, то есть на место, которое это значение занимает в ранжированном по возрастанию ряду выборочных данных. Следовательно, всегда $x_{(1)}$ – минимальное значение выборки, $x_{(n)}$ – её максимальное значение при объеме выборки n. Для значений x, попадающих в интервал между k-ым и $(k+1)$-ым значениями вариационного ряда, – $x_{(k)} < x \le x_{(k+1)}$, ЭФР $F_n(x) = k/n$. В частности, заметим, что если $x_{(k-1)} < x_{(k)}$, то $F_n(x_{(k)}) = (k-1)/n$.

В качестве примера рассмотрим ЭФР, построенную по 6 данным, среди которых -1 встречается два раза, 1 – три раза, а 4 – один раз. График ЭФР будет выглядеть следующим образом



Как и при построении гистограммы, график ЭФР полезно сравнить с графиком предполагаемого распределения (например, нормального). При этом некоторую информацию о степени достоверности этого распределения – правильности выдвинутого предположения о виде распределения – будет нести величина расхождения D, вычисленная по формуле (*). Неизвестные параметры модели можно оценить по выборке. В нормальной модели среднее μ оценивается выборочным средним $\overline{x}$, а дисперсия σ^2 – выборочной дисперсией s^2. Для показательного закона интенсивность отказа λ также может быть оценена посредством $\overline{x}$.

Если бы предполагаемое распределение было известно полностью и не надо было оценивать неизвестные параметры, то на основе значений D можно было бы построить критерий проверки адекватности этого распределения выборочным данным – так называемый критерий Смирнова (см., например, сборник таблиц [1]). Применять этот критерий к моделям с неизвестными параметрами нельзя, поскольку вероятность ошибки 1-го рода такого критерия будет зависеть от неизвестных параметров.

Глава III. Проверка гипотезы о типе распределения

Здесь описывается наиболее популярный метод проверки согласия выборочных данных с гипотезой о типе распределения.

Задание 4.

Критерий согласия хи-квадрат.

Постановка задачи.

Требуется проверить гипотезу $\mathbf{H}_0 : F \in \Psi$ о том, что функция распределения выборочных данных F принадлежит заданному семейству распределений Ψ (нормальному, экспоненциальному или равномерному).

Теоретические основы.

В качестве критерия проверки такой гипотезы чаще всего выбирают критерий согласия хи-квадрат. Для принятия решения в соответствии с этим критерием необходимо:

I) Выдвинуть гипотезу $\mathbf{H}_0 : F \in \Psi$ о виде распределения выборочных данных F.

II) Разбить область значений наблюдаемой с.в. на r интервалов.

III) По n выборочным данным подсчитать таблицу частот ν_i, аналогичную гистограммной таблице.

IV) Для каждого интервала вычислить теоретические вероятности p_i попадания в этот интервал.

V) Вычислить тестовую статистику

$$X^2 = \sum_{i=1}^{r} \frac{(\nu_i - n\,p_i)^2}{n\,p_i},$$

представляющую собой некую меру расхождения между ожидаемыми (теоретическими) частотами $n\,p_i$ и выборочными (полученными в эксперименте) частотами ν_i.

VI) Вычислить критический уровень значимости $\alpha_{\text{крит}}$.

VII) Принять решение о статистической значимости проверяемой гипотезы.

Разберем каждый пункт по порядку.

I) Чаще всего на практике возникает задача проверки гипотезы нормальности. В этом случае семейство Ψ есть семейство нормальных распределений $F(x) = \Phi\left(\dfrac{x-\mu}{\sigma}\right),\ \mu \in R^1, \sigma > 0,$ где

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-z^2/2} dz\ -$$

стандартное нормальное распределение, а (μ, σ) – два неизвестных параметра.

Если требуется проверить гипотезу о равномерном распределении на отрезке [0; 1], то семейство Ψ будет состоять всего из одного распределения $F_0(x) = x,\ 0 \le x \le 1.$

Если проверяется гипотеза об экспоненциальном типе распределения, то семейство Ψ состоит из функций вида $F(x) = 1 - \exp\left(-\dfrac{x}{\lambda}\right),\ x > 0,\ \lambda > 0.$

II) Строго говоря, разбиение числовой прямой на интервалы (как и число этих интервалов) необходимо выбирать независимо от наблюдённых в эксперименте значений. Однако практически это требование можно реализовать, только если гипотетическое распределение известно полностью. В противном случае очень часто будет наблюдаться ситуация, когда большинство выборочных данных попадут в один интервал. Поэтому на практике интервалы строят в соответствии с выборочными данными. В качестве одного из возможных способов такого построения предлагается просто разбить “*размах*” выборочных данных (от $x_{\min}$ до

$x_{\max}$) на r равных интервалов с последующим расширением двух крайних интервалов до $\pm\infty$.

Предостережение. Так делать нельзя. И уж, если очень хочется, то – можно. Имеется в виду, что так обычно и делают, но корректность статистического вывода в этом случае трудно будет обосновать.

С другой стороны, необходимо следить, чтобы выборочные значения не попадали на границу между интервалами. Это достигается путем задания значений границ интервалов с точностью на единицу большую, чем точность выборочных данных. При выполнении курсового проекта мы поступим проще. А именно, воспользуемся результатами задания 2.

III) Построение таблицы частот описано в задании на построение гистограммы.

IV) Вероятность попадания в i–ый интервал $[x_{i-1}; x_i)$ равна

$$p_i = F_0(x_i) - F_0(x_{i-1}),$$

где функция распределения F_0 либо совпадает с теоретической, если проверяется простая гипотеза, либо для её вычисления в гипотетическом распределении неизвестные параметры заменяются оценками. В качестве оценок параметров нормального распределения можно взять выборочное среднее $\tilde{\mu} = \bar{x}$ и выборочную дисперсию $\tilde{\sigma}^2 = s^2$. В этом случае

$$F_0(x) = \Phi\left(\frac{x - \bar{x}}{s}\right).$$

Для двух крайних интервалов $(-\infty; x_1)$ и $(x_{r-1}; \infty)$ вероятности равны $p_1 = F_0(x_1)$ и $p_r = 1 - F_0(x_{r-1})$, соответственно.

V) Процесс вычисления статистики хи-квадрат в среде `Excel` описан в пособии [4] на примере проверки гипотезы нормальности.

VI) Для вычисления критического уровня значимости необходимо знать функцию распределения $G(x)$ тестовой статистики X^2. Если бы гипотетическое распределение было известно точно, то есть были бы известны все параметры проверяемой модели (как при проверке гипотезы равномерности), то при большом объеме выборки функцию $G(x)$ можно

было бы аппроксимировать хи-квадрат распределением $K_{r\text{-}1}(x)$ с $(r\text{-}1)$-ой степенью свободы. В этом случае критический уровень значимости

$$\alpha_{\text{крит}} \approx 1 - K_{r-1}(x^2),$$

где x^2 – полученное значение статистики X^2. Если параметры модели оцениваются по выборке, то функция $G(x)$ начинает зависеть от этих параметров. Известно, однако, что если m неизвестных параметров оцениваются по методу максимального правдоподобия, то при $n \to \infty$ справедливо неравенство

$$K_{r-1}(x) \le G(x) \le K_{r-1-m}(x).$$

Поэтому, например, при проверке гипотезы нормальности (модель с двумя неизвестными параметрами) необходимо вычислить два значения:

$\alpha_{r\text{-}1} = 1 - K_{r\text{-}1}(x^2)$ и $\alpha_{r\text{-}3} = 1 - K_{r\text{-}3}(x^2)$. **Докажите**, что $\alpha_{r\text{-}3} < \alpha_{r\text{-}1}$.

Кстати, если границы интервалов выбирать в зависимости от данных, то поведение функции распределения $G(x)$ становится еще более непредсказуемым.

VII) Выводы о справедливости гипотезы делаются в зависимости от расположения $\alpha_{r\text{-}1\text{-}m}$, $\alpha_{r\text{-}1}$ относительно выбранного уровня значимости α. Если

- $\alpha_{r\text{-}1\text{-}m} > \alpha$ – гипотеза принимается с $\alpha_{\text{крит}} = \alpha_{r\text{-}1\text{-}m}$;
- $\alpha_{r\text{-}1} < \alpha$ – гипотеза отвергается с $\alpha_{\text{крит}} = \alpha_{r\text{-}1}$;
- $\alpha_{r\text{-}1\text{-}m} \le \alpha \le \alpha_{r\text{-}1}$ – нет достаточных оснований для принятия какого-либо решения ($\alpha_{\text{крит}} = \alpha$).

Замечание. Критерий хи-квадрат носит название критерия согласия, поскольку при его применении не учитывается вид альтернативы и вывод, который при этом делается, означает только, что или мы гипотезу принимаем – данные согласуются с гипотезой, или отвергаем – данные не согласуются с гипотезой.

Глава IV. Проверка гипотезы однородности

В практике применения методов статистического анализа наиболее востребованы методы сравнения различных совокупностей выборочных данных с целью выяснения их однородности. Например, при исследовании лечебных свойств нового препарата часто требуется

а) сравнить воздействие этого препарата на некоторую характеристику здоровья в одной группе пациентов (– первая выборка) с воздействием старого препарата на ту же характеристику в другой группе пациентов (– вторая выборка), или

б) сравнить характеристики здоровья у пациентов одной группы до лечения (– первая выборка) и после лечения препаратом (– вторая выборка), или

в) сравнить долю выздоровевших пациентов при различных способах лечения.

В общем виде задачу можно сформулировать следующим образом.

Имеются две совокупности выборочных данных. Требуется сравнить их между собой.

Случайные выборки можно сравнить только по их функциям распределения. Таким образом, совпадение совокупностей означает, что они имеют одинаковое распределение или, другими словами, однородны.

Здесь мы изучим 6 критериев сравнения совокупностей. На практике выбор того или иного критерия зависит от ответа на следующие два основных вопроса.

1) Имеют ли выборки нормальное распределение?
2) Можно ли считать наблюдения в различных выборках независимыми?

Задание 5.

Одновыборочный критерий Стьюдента.

Постановка задачи.

Двухвыборочный вариант. Имеются две выборки $(x_1,...,x_n),(y_1,...,y_n)$ одинакового объема. Известно, что распределения в этих выборках подчинены нормальному закону и, кроме того, каждое i-ое наблюдение x_i в 1-ой выборке зависит (в вероятностном смысле) от соответствующего i-ого наблюдения y_i во второй выборке. Требуется проверить гипотезу однородности выборок. Точнее, требуется проверить гипотезу о том, что среднее значение разности выборок равно нулю (меньше нуля, больше нуля).

Одновыборочный вариант. Имеется одна выборка из нормального распределения. Требуется проверить гипотезу о том, что среднее значение этого распределения не превосходит заданной величины $C_{\text{норм}}$.

Теоретические основы.

Поскольку каждое значение x_i в одной выборке зависит от значения y_i в другой выборке (например, оба значения суть измерения одной и той же характеристики у одного пациента), то вместо двух измерений x_i, y_i рассматривают их разность $u_i = x_i - y_i$. Таким образом, если обе выборки получены из нормального распределения, то гипотеза однородности выборок может быть сформулирована в виде

$$\mathbf{H}_0 : \mu = 0,$$

где $\mu = \mu_1 - \mu_2$ – разность средних значений первой и второй выборок. Альтернатива к нулевой гипотезе чаще всего совпадает с ожиданиями экспериментатора. Например, если выборки представляют собой измерения верхнего артериального давления до и после лечения новым препаратом от гипертонии, то естественно ожидать, что во второй выборке значения будут ниже, чем в первой. В этом случае альтернатива должна

иметь вид $\mathbf{H}_1 : \mu_1 - \mu_2 > 0$. С другой стороны, если экспериментатору важен лишь сам факт отличия одной выборки от другой, то следует выбирать двухстороннюю альтернативу вида $\mathbf{H}_1 : \mu_1 - \mu_2 \neq 0$.

Статистика одновыборочного критерия Стьюдента равна

$$T = \frac{\overline{u}}{s_u}\sqrt{n-1},$$

где $\overline{u}$ – выборочное среднее, s_u – выборочная дисперсия (смещенная оценка), вычисленные по разностям $u_i = x_i - y_i$, n – объем выборки. Если справедливо предположение о нормальности распределения и верна нулевая гипотеза, то статистика T имеет распределение Стьюдента $S_{n-1}(t)$ с $(n-1)$-ой степенью свободы (см.введение). В любом учебнике по математической статистике имеются таблицы этого распределения. Следует помнить, что распределение Стьюдента симметрично – $S_{n-1}(-t) = 1 - S_{n-1}(t)$, поэтому таблицы построены обычно лишь для входных значений $t > 0$. Таким образом, если t – результат вычислений статистики T по экспериментальным данным, то критический уровень значимости равен

альтернатива	критический уровень значимости	пояснения
$\mathbf{H}_1 : \mu_1 - \mu_2 > 0$	$\alpha_{\text{крит}} = 1 - S_{n-1}(t)$	$= \mathbf{P}\{T > t\}$
$\mathbf{H}_1 : \mu_1 - \mu_2 < 0$	$\alpha_{\text{крит}} = S_{n-1}(t)$	$= \mathbf{P}\{T < t\}$
$\mathbf{H}_1 : \mu_1 - \mu_2 \neq 0$	$\alpha_{\text{крит}} = 2(1 - S_{n-1}(t))$	$= \mathbf{P}\{\lvert T \rvert > t\}$

Из опыта применения этого критерия замечено, что очень часто при вычислении разностей переставляются выборки (вместо $x_i - y_i$ вычисляется $y_i - x_i$), что не является криминалом. Необходимо лишь следить за адекватностью выдвинутой односторонней альтернативы и избранного способа вычисления статистики Стьюдента. Для двухсторонней альтернативы способ вычисления не принципиален.

Замечание. Критерий потому и называется одновыборочным, что по своей сути предназначен для сравнения среднего значения одной нормальной выборки с некоторой нормой $C_{\text{норм}}$. Описанную выше схему с

очевидными изменениями можно применить и в этом случае. Например, если перед исследователем стояла задача полного излечения гипертонических больных, то необходимо было бы проверить гипотезу о том, что среднее значение верхнего артериального давления у пациентов, прошедших курс лечения, будет больше 125 при альтернативе меньше 125. Отличие заключается только в том, что вместо разностей $u_i = x_i - y_i$ следует рассмотреть разности $u_i = x_i - C_{\text{норм}}$.

Задание 6.

Критерий знаков.

Постановка задачи.

Двухвыборочный вариант. Имеются две выборки $(x_1,..., x_n), (y_1,..., y_n)$ одинакового объема. Известно, что каждое i-ое наблюдение x_i в 1-ой выборке зависит (в вероятностном смысле) от соответствующего i-ого наблюдения y_i во второй выборке. Распределение выборок неизвестно. Требуется проверить гипотезу однородности выборок.

Одновыборочный вариант. Имеется одна выборка. Требуется проверить гипотезу, что некоторое фиксированное событие происходит чаще, чем противоположное к этому событию утверждение (например, лечение чаще приводит к выздоровлению).

Теоретические основы.

Если исследователя интересует лишь факт наличия эффекта воздействия и нет оснований предполагать нормальность распределения выборок, то можно каждую пару исходных данных (x_i, y_i) заменить величиной z_i, принимающей всего два значения: $z_i = 1$, если эффект есть, и $z_i = 0$, если эффекта нет. Под эффектом может пониматься, например, уменьшение артериального давления после лечения или увеличение доли полезных веществ после стерилизации. Если предположить, что воздействие не обладает никаким эффектом, то величина z_i будет иметь распределение Бернулли с вероятностью успеха p = 0,5 (– приблизительно в 50% случаев

должен наблюдаться эффект). Исследователю было бы интересно проверить гипотезу

$\mathbf{H}_0 : p \le 0{,}5$ – эффект отсутствует или направлен противоположно,

при альтернативе $\mathbf{H}_1 : p > 0{,}5$.

Для проверки гипотез подобного рода можно использовать критерий знаков, статистика которого M равна числу благоприятных исходов:

$$M = \sum_{i=1}^{n} z_i .$$

Из определения видно, что M имеет биномиальное распределение с параметрами (n, p). Нулевая гипотеза должна отвергаться, если выборочное значение m статистики M будет достаточно большим. Критический уровень значимости такого критерия равен (по формуле для биномиального распределения)

$$\alpha_{\text{крит}} = \sup_{p \le 0{,}5} \mathbf{P}\{M \ge m \mid n, p\} = \sup_{p \le 0{,}5} \sum_{k=m}^{n} C_n^k p^k (1-p)^{n-k} .$$

Нетрудно понять, что supremum в последнем выражении достигается на границе между гипотезами при p = 0,5 (чем больше p, тем больше ожидаемое значение M). Таким образом, при выборочном значении $M = m$

$$\alpha_{\text{крит}} = \frac{1}{2^n} \sum_{k=m}^{n} C_n^k . \qquad (*)$$

Замечание. Описанную схему можно применять также для проверки гипотезы о вероятности “успеха” при биномиальных испытаниях – одновыборочный вариант критерия. В качестве примера рассмотрим ситуацию, когда при составлении договора купли-продажи заказчиком была оговорена нижняя граница в 92% для доли доброкачественной продукции. При поступлении товара заказчик проводит контрольные измерения n единиц продукции. По результатам испытаний, основываясь только на количестве кондиционной продукции, требуется проверить гипотезу $\mathbf{H}_0 : p \le 0{,}92$ (опять же гипотеза противоположна ожиданиям).

Другой пример. С целью прогнозирования результатов будущих выборов было опрошено 1000 респондентов. Среди них оказалось 35

сторонников партии «Будет ещё хуже!». Можно ли утверждать, что эта партия не попадет в думу?

Если считать 1000 респондентов каплей в море всех избирателей и отбор респондентов производился абсолютно случайно, то можно описать наши наблюдения как выборку из распределения Бернулли с вероятностью успеха p, равной доле всех сторонников указанной партии. Если мы находимся на позициях противников партии «Будет ещё хуже!», мы хотели бы, что бы эта доля была меньше 0,05. Поэтому в качестве альтернативы мы должны выбрать утверждение $\mathbf{H}_1 : p \le 0{,}05$.

Итак, $n = 1000, m = 35$, граничное значение $p_0 = 0{,}05$. Критический уровень значимости равен $\alpha_{\text{крит}} = \mathbf{P}\{M \le m \mid p = 0{,}05\} = 0{,}014$.

Вывод. Скорее всего (с надежностью 98,6%), партия «Будет ещё хуже!» не пройдет в Думу.

Задание 7.

Двухвыборочный критерий Стьюдента.

Постановка задачи.

Имеются две выборки $(x_1,\ldots,x_{n_1}), (y_1,\ldots,y_{n_2})$, относящиеся к двум независимым группам наблюдений одной и той же характеристики, подчиняющейся нормальному закону с одинаковыми для обеих выборок дисперсиями. Требуется проверить гипотезу однородности выборок.

Теоретические основы.

Однородность выборок в предположении нормальности их распределения эквивалентна совпадению средних и дисперсий. Критерий Стьюдента применяется в том случае, если можно априори, по тем или иным соображениям, предположить, что дисперсии одинаковы. Например, одна и та же характеристика измеряется одним и тем же прибором, и разброс данных обусловлен исключительно ошибками этого прибора. Другой пример, обычный для медицинской практики, – сравнение показателей здоровья в двух группах пациентов, подвергнутых двум различным

методам лечения, причем группы сформированы одинаковым образом. В этом случае гипотеза однородности эквивалентна равенству средних $\mathbf{H}_0 : \mu_1 = \mu_2$.

Статистика двухвыборочного критерия Стьюдента равна

$$T = \frac{\bar{x} - \bar{y}}{\sqrt{n_1 s_x^2 + n_2 s_y^2}} \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}},$$

где $\bar{x}, \bar{y}$ – выборочные средние, а s_x^2, s_y^2 – выборочные дисперсии (смещенные оценки) первой и второй выборки, соответственно. Если обе выборки имеют нормальное распределение и верна нулевая гипотеза, то статистика Стьюдента T имеет распределение Стьюдента $S_{n_1+n_2-2}(t)$ с $(n_1 + n_2 - 2)$ степенями свободы.

Поэтому критический уровень значимости при односторонней альтернативе $\mathbf{H}_1 : \mu_1 < \mu_2$ равен $\alpha_{\text{крит}} = \mathbf{P}\{T \le t\} = S_{n_1+n_2-2}(t)$.

Задание. Запишите критический уровень значимости для противоположной альтернативы и двухсторонней альтернативы.

Задание 8.

Критерий Вилкоксона.

Постановка задачи.

Имеются две выборки $(x_1,...,x_{n_1}), (y_1,...,y_{n_2})$, относящиеся к двум независимым группам наблюдений одной и той же характеристики. Требуется проверить гипотезу однородности выборок в ситуации, когда ожидается, что значения в 1-й выборке будут меньше значений во второй выборке.

Теоретические основы.

Описываемый здесь критерий применяется в том случае, когда

а) распределение выборки не подчиняется нормальному закону и

б) в качестве альтернативы однородности выборок выдвигается гипотеза

$$\mathbf{H}_1 : F_Y(x) = F_X(x - \Delta), \Delta > 0.$$

Другими словами, распределение первой выборки (*x*-ов) сдвинуто влево относительно распределения второй выборки (*y*-ов), то есть ожидаемые значения *x*-ов должны быть меньше значений *y*-ов. Идею критерия Вилкоксона можно проиллюстрировать на следующем “крайнем” примере. Предположим, что все выборочные значения из 2-ой выборки больше всех значений из 1-ой выборки. Такая ситуация вполне ожидаема, если верна альтернатива. В этом случае, расположив обе выборки в один общий ряд, мы видим, что 1-ая выборка занимает меньшие по порядку места (ранги), чем 2-ая выборка. Если же обе выборки равномерно перемешаны (что естественно, если верна гипотеза), то средние ранги для обеих выборок должны быть приблизительно равны. Таким образом, малые значения средних рангов 1-ой выборки будут свидетельствовать в пользу альтернативы.

Для построения критерия Вилкоксона необходимо обе выборки расположить в один общий ряд, упорядоченный по возрастанию. При этом информация о принадлежности каждого значения к той или иной выборке не должна быть утеряна. Статистика Вилкоксона равна

$$W = \sum_{i=1}^{n_2} r_i ,$$

где $r_1, \ldots r_{n_2}$ – ранги всех значений 1-ой выборки (– той, для которой альтернатива предполагает сдвиг влево). Нулевая гипотеза отвергается, если для полученного в эксперименте значения $W = w$

$$w < C_{\text{крит}} .$$

Критическая константа $C_{\text{крит}}$, как всегда, находится из условия

$$\mathbf{P}_0\{W < C_{\text{крит}}\} \le \alpha ,$$

где вероятность $\mathbf{P}_0$ вычислена в предположении, что верна нулевая гипотеза. Конкретное значение $C_{\text{крит}}$ для конкретных значений объемов выборок n_1 и n_2 и уровня значимости α можно найти в таблицах математической статистики (см.сборник таблиц [1]). Например, при $n_1 = 8$, $n_2 = 10$ и $\alpha = 0{,}05$ критическая константа $C_{\text{крит}} = 56$.

Указанная таблица построена таким образом, что объем 1-ой выборки не должен быть больше объема 2-ой выборки. Если это не

выполняется, то критическую константу следует изменить на $C_{\text{крит}} + (n_1 - n_2)(n_1 + n_2 + 1)/2$. Таким образом, при $n_1 = 10$, $n_2 = 8$ и $\alpha = 0{,}05$ максимальное значение статистики Вилкоксона, при которой гипотеза однородности будет отвергаться равно,

$C_{\text{крит}} = 75$ (=56 + (10-8)*(10+8+1)/2).

Другая форма критерия, как обычно, связана с критическим уровнем значимости

$$\alpha_{\text{крит}} = \mathbf{P}_0\{W \le w\}.$$

К сожалению, таблицы распределения W труднодоступны. Поэтому единственный путь нахождения $\alpha_{\text{крит}}$ – воспользоваться асимптотическим (при $n_1, n_2 \to \infty$) распределением статистики W. Известно, что статистика W асимптотически нормальна

со средним $\mu_W = n_1(n_1 + n_2 + 1)/2 + 0{,}5$

и дисперсией $\sigma_W^2 = n_1 n_2 (n_1 + n_2 + 1)/12$.

То есть, если w – выборочное значение статистики W, полученное по рангам 1-ой выборки, то для альтернативы $\mathbf{H}_1$: «1-ая выборка сдвинута влево», – критический уровень значимости равен

$$\alpha_{\text{крит}} = \mathbf{P}_0\{W \le w\} \approx \Phi\left(\frac{w - \mu_W}{\sqrt{\sigma_W^2}}\right),$$

где Ф – стандартная нормальная функция распределения.

Другая проблема, связанная с построением критерия Вилкоксона – совпадающие наблюдения. Если два наблюдения имеют одинаковые значения и принадлежат к одной группе, то статистика Вилкоксона не будет изменяться при случайных перестановках этих наблюдений в общем ряду данных. Если же совпадающие наблюдения принадлежат разным группам, то встает вопрос, какое из этих наблюдений следует поставить раньше? Чтобы избежать ненужного здесь “волюнтаризма”, можно всем совпадающим значениям присвоить один и тот же ранг, равный среднему арифметическому мест, на которых эти значения находятся. Например, если четыре совпадающих значения занимают места с 9-го по 12-е, то все они получают ранг 10,5. Распределение измененной таким образом статистики Вилкоксона очень тяжело вычислить. В качестве “первого

приближения” можно воспользоваться описанной выше методикой нахождения критической константы по таблицам или методикой вычисления асимптотического уровня значимости.

Заметим, как удивительно точно “работает” асимптотическая формула для критического уровня значимости. Приведем фрагмент таблицы Большева Л.Н., Смирнова Н.В. [1] с точными значениями критической константы для критерия Вилкоксона:

n_1	n_2	α					
		0,001	0,005	0,01	0,025	0,05	0,10
8	10	42	47	49	53	56	60

Если воспользоваться этой таблицей для обратной задачи нахождения критического уровня значимости по значению статистики Вилкоксона, то легко заметить, что для $w = 54$ точная величина критического уровня значимости находится в интервале (0,025; 0,05), причем явно ближе к его левому краю. Асимптотическое значение равно 0,028.

Задание 9.

Критерий Фишера. Критерий сравнения дисперсий.

Постановка задачи.

Имеются две выборки $(x_1,...,x_{n_1}), (y_1,...,y_{n_2})$, относящиеся к двум независимым группам наблюдений одной и той же характеристики, подчиняющейся нормальному закону. Требуется сравнить дисперсии наблюдений в этих групп.

Теоретические основы.

Чаще всего требуется выяснить, больше или меньше дисперсия 1-ой совокупности по сравнению с дисперсией 2-ой совокупности. В этом случае нулевая гипотеза будет иметь вид

$$\mathbf{H}_0 : \sigma_x^2 = \sigma_y^2 \quad \text{или} \quad \mathbf{H}_0 : \frac{\sigma_x^2}{\sigma_y^2} = 1.$$

Для проверки этой гипотезы применяют критерий Фишера, тестовая статистика которого равна

$$f = \frac{s_x^2/(n_1 - 1)}{s_y^2/(n_2 - 1)}.$$

В предположении нормальности данных и при справедливости нулевой гипотезы статистика f имеет распределение Фишера $F_{k,m}(x)$ с параметрами $k = n_1 - 1$ и $m = n_2 - 1$. Таблицы этого распределения имеются в большинстве справочников по математической статистике. Таким образом, критический уровень значимости критерия Фишера равен:

альтернатива	уровень значимости
$\mathbf{H}_1 : \sigma_x^2 / \sigma_y^2 > 1$	$\alpha_{\text{крит}} = 1 - F_{k,m}(f)$
$\mathbf{H}_1 : \sigma_x^2 / \sigma_y^2 < 1$	$\alpha_{\text{крит}} = F_{k,m}(f)$
$\mathbf{H}_1 : \sigma_x^2 / \sigma_y^2 \neq 1$	$\alpha_{\text{крит}} = \begin{cases} 2(1 - F_{k,m}(f)), \text{ если } s_x^2 \geq s_y^2, \\ 2F_{k,m}(f), \text{ если } s_x^2 < s_y^2. \end{cases}$

Замечание. Во многих учебниках по математической статистике рекомендуют предварить применение двухвыборочного критерия Стьюдента проверкой гипотезы о равенстве дисперсий в группах. В соответствии с такой рекомендацией две выборки будут считаться не однородными, если или критерий Фишера отвергнет равенство дисперсий (обозначим такое событие через F), или критерий Фишера признает дисперсии равными (событие F^c), но критерий Стьюдента отвергнет гипотезу равенства средних значений (событие S). Другими словами, гипотеза однородности будет отвергаться, если произойдет событие $F + F^c S$. Размер такого критерия равен вероятности

$$\mathbf{P}\{F + F^c S\} = \mathbf{P}\{F\} + \mathbf{P}\{F^c S\}.$$

Если размер критерия Фишера равен α_1, а критерия Стьюдента – α_2, то первое слагаемое в последнем равенстве равно α_1. Про второе

слагаемое можно сказать только, что оно не больше α_2. Таким образом, уровень значимости составного критерия $\leq \alpha_1 + \alpha_2$. Достичь желаемого уровня в 5% (вернее, меньше, чем 5%) в этом случае можно, если положить $\alpha_1 = \alpha_2 = 0.025$.

Задание 10.

Критерий однородности хи-квадрат.

Постановка задачи.

Имеются две выборки $(x_1,...,x_{n_1}), (y_1,...,y_{n_2})$, относящиеся к двум независимым группам наблюдений одной и той же характеристики. Требуется проверить гипотезу однородности выборок в ситуации, когда неизвестна модель распределения выборок и нет никакой информации о соотношении между этими выборками.

Теоретические основы.

Рассмотрим, наконец, ситуацию, когда нет никакой информации ни о нормальности распределения данных, ни о соотношении между группами типа “левее - правее”. В этом случае можно воспользоваться идеей гистограммного представления данных и сравнить частоты попадания результатов измерений в различных группах в одни и те же интервалы числовой прямой.

Итак, пусть имеются две группы измерений объемов n_1 и n_2, соответственно, для которых подсчитаны частоты $\nu_{i1}, i = 1,...,r,$ и $\nu_{i2}, i = 1,...,r,$ попадания данных в r одинаковых интервалов. Если гипотеза однородности выборок верна, то относительные частоты ν_{i1}/n_1 и ν_{i2}/n_2 должны быть близки друг к другу. Это соображение приводит нас к следующей тестовой статистике.

Для каждого из интервалов $i = 1,...,r,$ подсчитаем общее число данных $\nu_{i\bullet} = \nu_{i1} + \nu_{i2}$, попавших в этот интервал. Статистика критерия однородности хи-квадрат равна

$$X^2 = n_1 n_2 \sum_{i=1}^{r} \frac{1}{\nu_{1\bullet}} \left(\frac{\nu_{i1}}{n_1} - \frac{\nu_{i2}}{n_2} \right)^2 .$$

При справедливости гипотезы однородности распределение статистики X^2 можно аппроксимировать распределением хи-квадрат с (r-1)-ой степенью свободы:

$$\mathbf{P}\{X^2 < x\} \approx K_{r-1}(x) \quad (n_1, n_2 \to \infty).$$

Ясно, что при справедливости гипотезы однородности статистика X^2 будет принимать “малые” значения. Поэтому, если x^2 – подсчитанное по экспериментальным данным значение статистики X^2, то критический уровень значимости критерия хи-квадрат будет равняться

$$\alpha_{\text{крит}} = \mathbf{P}\{X^2 > x^2\} \approx 1 - K_{r-1}(x^2).$$

Замечание 1. Кроме вывода об однородности или неоднородности групп, здесь полезно визуально сравнить распределения в группах. Для этого можно совместить гистограммы обеих выборок. Следует только помнить, что, поскольку объемы выборок могут быть различны, то гистограммы должны быть построены по относительным (деленным на объемы выборок) частотам.

Замечание 2. Построенный критерий не зависит от способа, каким были получены частоты. Этот критерий можно использовать и для проверки однородности двух выборок, когда частоты представляют собой количества выборочных данных, удовлетворяющих произвольным взаимоисключающим условиям. Например, в пособии “Курсовой проект …” [2] рассматривается задача сравнения двух общин по группам крови (r = 4). Другой пример. В медицинской практике очень часто требуется сравнить новую методику лечения со старой методикой по результатам клинических наблюдений. При этом пациентов, прошедших курс лечения подразделяют, например, на 3 группы – а) не выздоровели, б) выздоровели, но через год болезнь повторилась, и в) выздоровели без последующего рецидива.

Больше двух выборок.

Критерий однородности хи-квадрат может быть применен и в случае, если число выборок больше двух. Пусть ν_{ij} – число исходов в j-ой выборке ($j=1,...,s$), попавших в i-ый интервал группировки ($i=1,...,r$). Таким образом, данные могут быть представлены в виде таблицы, где , как и раньше, точка на месте одного из индексов означает суммирование данных по этому индексу при фиксированном другом индексе. Так, например, общий объем выборок равен $n=\nu_{\bullet\bullet}$.

Выборка / Интервал	1	…	s	Всего
1	ν_{11}	…	ν_{1s}	$\nu_{1\bullet}$
⋮	…			⋮
r	ν_{r1}	…	ν_{rs}	$\nu_{r\bullet}$
Всего	$\nu_{\bullet 1}$	…	$\nu_{\bullet s}$	$n=\nu_{\bullet\bullet}$

В предположениях гипотезы однородности статистика

$$X^2 = n\left(\sum_{j=1}^{s}\sum_{i=1}^{r}\frac{\nu_{ij}^2}{\nu_{i\bullet}\nu_{\bullet j}} - 1\right)$$

имеет асимптотическое хи-квадрат распределение с $\nu=(r-1)(s-1)$ степенями свободы:

$$\mathbf{P}\{X^2 < x\} \approx K_\nu(x), \quad n\to\infty.$$

Поэтому гипотеза однородности всех s выборок должна отвергаться, если критический уровень значимости

$$\alpha_{\text{крит}} \approx 1 - K_\nu(x^2)$$

меньше выбранного уровня значимости α.

Глава V. Интервальные оценки

Теоретические основы.

Пусть $x^{(n)} = (x_1, \ldots, x_n)$ – случайная выборка, распределение которой $F_\theta(x)$ зависит от некоторого неизвестного параметра θ. Интервал $(\underline{\theta}; \overline{\theta})$ с границами $(\underline{\theta}(x^{(n)}); \overline{\theta}\ (x^{(n)}))$, зависящими от выборочных данных, называется

(1-α)-доверительным интервалом для параметра θ, если

$$\mathbf{P}_\theta\{\underline{\theta} < \theta < \overline{\theta}\} \geq 1-\alpha. \qquad (*)$$

Статистика $\overline{\theta}$ называется

верхней (1-α)-доверительной границей для параметра θ, если

$$\mathbf{P}_\theta\{\theta < \overline{\theta}\} \geq 1-\alpha.$$

Статистика $\overline{\theta}$ называется

нижней (1-α)-доверительной границей для параметра θ, если

$$\mathbf{P}_\theta\{\underline{\theta} < \theta\} \geq 1-\alpha.$$

Интерпретация.

Смысл этих определений легко понять, если вспомнить, что индекс θ, стоящий у знака вероятности $\mathbf{P}_\theta$, указывает на истинное значение неизвестного параметра. Поэтому формула (*), например, означает, что с большой вероятностью доверительный интервал накроет истинное значение оцениваемого параметра. На практике обычно делается несколько вольный вывод, что с большой долей вероятности следует ожидать значение оцениваемого параметра, принадлежащее интервалу $(\underline{\theta}; \overline{\theta})$. В таком утверждении “скрытно” присутствует предположение о случайности изменения параметра θ от эксперимента к эксперименту. В действительности, оцениваемый параметр не случаен, а имеет некоторое фиксированное неизвестное значение.

Точность и надежность интервала.

Величина $Q = (1-\alpha)\cdot 100\%$ называется *надежностью* интервала и выбирается обычно в пределах от 90% до 99% (стандартное значение – 95%). На первый взгляд кажется, что чем выше значение надежности, тем лучше будет построенный интервал. Однако здесь надо учитывать, что чем больше величина Q, тем шире получится доверительный интервал (в пределе при $\alpha = 0$ он будет совпадать с R^1), то есть уменьшится его точность. Задача построения доверительного интервала с заданной точностью и надежностью может быть решена только при достаточно большом объеме выборки.

Двухсторонний интервал через доверительные границы.

Для построения $(1-\alpha)$-доверительного интервала $(\underline{\theta};\overline{\theta})$ можно построить отдельно верхнюю $\overline{\theta}$ и нижнюю $\underline{\theta}$ границы с надежностью $(1-\alpha/2)\cdot 100\%$.

Связь с задачей проверки гипотез.

Пусть $B(x^{(n)})$ – некое $(1-\alpha)$-доверительное множество. Тогда критерий, отвергающий гипотезу, если $B(x^{(n)})$ полностью попадает в область альтернативы, будет иметь уровень α. Так, при альтернативе $\mathbf{H}_1 : \theta > \theta_0$ гипотезу следует отвергать, если нижняя граница $\underline{\theta} > \theta_0$. Если же ошибочно принимать гипотезу, когда доверительное множество полностью попадает в область гипотезы (например, верхняя граница $\overline{\theta} \le \theta_0$), то такой критерий будет иметь вовсе “неприемлемый” уровень $1-\alpha$, вместо ожидаемого уровня α.

Обратно, рассмотрим задачу проверки простой гипотезы $\mathbf{H}_\vartheta : \theta = \vartheta$ о параметре θ распределения наблюдаемой случайной величины. Пусть $A(\vartheta; x^{(n)})$ – критическая область уровня α (область, где гипотеза отвергается). Тогда множество $B(x^{(n)})$ тех значений параметра ϑ, при которых гипотеза H_ϑ принимается, –

$$B(x^{(n)}) = \{\vartheta : x^{(n)} \notin A(\vartheta; x^{(n)})\},$$

образует (1-α)-доверительное множество. Множество $B(x^{(n)})$ определяет

a) нижнюю границу, если альтернатива имеет вид $\mathrm{K}: \theta > \vartheta$;
b) верхнюю границу, если альтернатива имеет вид $\mathrm{K}: \theta < \vartheta$;
c) двухсторонний интервал, для альтернативы вида $\mathrm{K}: \theta \neq \vartheta$.

Задание.

1) Проинтерпретируйте смысл определений односторонних доверительных границ.
2) Докажите, что критерий, построенный по $(1-\alpha)$–доверительному множеству, имеет уровень α.

Методы построения.

I. *Метод опорной функции*. Пусть для некоторой статистики $T = T(x^{(n)})$ существует монотонное по параметру θ преобразование G (t, θ) (так называемая *опорная функция*), для которого функция распределения $F(x) = \mathbf{P}_\theta\{G(T,\theta) < x\}$ не зависит от θ. Тогда, выбирая Δ из соотношения $F(\Delta) = 1-\alpha$ и разрешая неравенство $G(t,\theta) < \Delta$ относительно θ при полученном экспериментальном значении статистики $T = t$, получаем верхнюю или нижнюю (в зависимости от направления монотонности G) доверительную границу для θ.

Этот метод применяется при построении доверительных границ для среднего значения и дисперсии нормального распределения.

II. *Метод, основанный на функции распределения оценки.* Пусть распределение $F(t,\theta) = \mathbf{P}_\theta(T < t)$ статистики $T = T(x^{(n)})$ непрерывно убывает с ростом параметра θ. Тогда, если экспериментальное значение статистики T=t, то значение нижней (1-α)-доверительной границы $\underline{\theta}$ можно получить как решение уравнения $F(t,\underline{\theta}) = 1-\alpha$. Верхняя (1-$\alpha$)-доверительная граница получается как решение уравнения $\mathbf{P}_\theta\{T > t\} = 1-\alpha$.

Этот метод применяется при построении доверительных границ для вероятности наблюдаемого события.

II. *Метод, основанный на асимптотическом распределении оценок.* Этот метод близок к первому методу. Если известно асимптотическое распределение некоторой статистики T, то, оценив мешающие параметры, можно построить опорную функцию, предельное распределение которой не будет зависеть от неизвестных параметров. Далее, поступая как и в методе I, можно построить доверительную границу для неизвестного параметра. Надежность такого доверительного утверждения с ростом объема выборки будет приближаться к номинальной надежности Q.

Этот метод также применяется при построении доверительных границ для вероятности наблюдаемого события.

Задание 11.

Построить интервальную оценку для среднего значения нормального распределения.

Постановка задачи.

Имеется выборка $(x_1,...,x_n)$ из нормального распределения. Требуется построить 95%-доверительный интервал (верхнюю границу, нижнюю границу) для неизвестного среднего μ этого распределения.

Теоретические основы.

Пусть $\bar{x}$ - выборочное среднее, s^2 - выборочная дисперсия, вычисленные по выборке из нормального распределения со средним μ и дисперсией σ^2. Опорная функция

$$G = \frac{\overline{X} - \mu}{S}\sqrt{n-1}$$

монотонно убывает по μ и имеет распределение Стьюдента $S_{n-1}(t)$ с (n-1)-ой степенью свободы (см. введение). Пусть $t^\alpha = t^\alpha(n-1) = S_{n-1}^{-1}(1-\alpha)$ -

верхняя α-квантиль распределения $S_{n-1}(t)$ (то есть, решение уравнения $S_{n-1}(t)=1-\alpha$), тогда с надежностью (1-α)·100%

а) $\underline{\mu}=\overline{x}-\frac{s}{\sqrt{n-1}}t^{\alpha}$ - нижняя доверительная граница для среднего;

б) $\overline{\mu}=\overline{x}+\frac{s}{\sqrt{n-1}}t^{\alpha}$ - верхняя доверительная граница для среднего.

Как видно из формул для доверительного интервала, ширина этого интервала пропорциональна отношению $s/\sqrt{n-1}$. Это отношение называется *стандартной ошибкой среднего*, обозначается обычно буквой m и весьма оригинально читается – "эм малое". В медицинской практике принято результаты вычислений записывать в виде $4{,}366\pm0{,}107$, где первое слагаемое есть среднее арифметическое $\overline{x}$, а второе слагаемое – ошибка среднего m.

Квантиль распределения t^{α} чаще всего находят по таблицам. Во введении мы нашли, что при объеме выборки $n=20$ верхняя 95%-ая квантиль распределения Стьюдента равна $t^{0,05}(19)=1{,}7291$. Эту константу применяют для построения любой односторонней 95%-доверительной границы при 20 наблюдениях. Для построения двухстороннего 95%-доверительного интервала применяется константа $t^{0,025}(19)=2{,}0930$.

Задание 12.

Построить интервальную оценку для дисперсии нормального распределения.

Постановка задачи.

Имеется выборка $(x_1,...,x_n)$ из нормального распределения. Требуется построить 95%-доверительный интервал (верхнюю границу, нижнюю границу) для неизвестной дисперсии σ^2 этого распределения.

Теоретические основы.

Пусть, как и выше, s^2 – выборочная дисперсия, построенная по выборке из нормального распределения с неизвестной дисперсией σ^2, тогда опорная функция $G = n\frac{S^2}{\sigma^2}$ имеет распределение хи-квадрат $K_{n-1}(x)$ с (n–1)-й степенью свободы (см. введение). Таким образом,

а) $\underline{\sigma}^2 = \dfrac{n s^2}{t^{\alpha}(n-1)}$ – нижняя (1-α)-доверительная граница;

б) $\overline{\sigma}^2 = \dfrac{n s^2}{t^{1-\alpha}(n-1)}$ – верхняя (1-α)-доверительная граница,

где $t^p(n-1) = K_{n-1}^{-1}(1-p)$ – верхняя квантиль распределения хи-квадрат, которая может быть найдена по таблицам (см. Глава I, стр. 19). Таким образом, при 19 степенях свободы $t^{0,025}(19) = 32{,}852$, $t^{0,975}(19) = 8{,}907$.

Задание 13.

Построить интервальную оценку для вероятности успеха

Постановка задачи.

В эксперименте подсчитывалось число успешных реализаций некоторого события (например, число доброкачественных изделий). Требуется построить доверительную границу для вероятности p этого события.

Теоретические основы.

Мы начнем с наиболее простого асимптотического метода построения границ (метод III). Обозначим через T случайную величину, равную числу успехов в n независимых наблюдениях. По известной теореме Муавра-Лапласа статистика T в пределе при $n \to \infty$ имеет нормальное распределение со средним np и дисперсией $np(1-p)$. Таким образом,

$$\mathbf{P}_\theta\left\{\frac{T-np}{\sqrt{np(1-p)}}<x\right\}\to\Phi(x).$$

Положив в последнем соотношении $x=x^\alpha=\Phi^{-1}(1-\alpha)$ и разрешив неравенство под знаком вероятности относительно p, получим нижнюю границу для p. Такой способ слишком сложен (надо владеть школьными методами решения неравенств с радикалами ☺), поэтому мы оставим его в качестве самостоятельного задания, а сами пойдем по более простому пути. А именно: в выражении для асимптотической дисперсии значение неизвестного параметра p оценим величиной $\tilde{p}=t/n$, где t – наблюдённое в эксперименте число успешных реализаций события. Тогда можно утверждать, что с вероятностью, близкой к $1-\alpha$, будет выполняться неравенство

$$p>\tilde{p}-\frac{\sqrt{\tilde{p}(1-\tilde{p})}}{\sqrt{n}}\cdot t^\alpha.$$

Правая часть этого неравенства дает нижнюю границу для p. Кстати, заметим, что сомножитель $\sqrt{\tilde{p}(1-\tilde{p})/n}$ также называют *стандартной ошибкой*. Очевидно, верхняя граница получится, если знак "–" перед стандартной ошибкой заменить на "+".

Надежность только что построенных границ при малых объемах выборки вызывает некоторое сомнение. К счастью, в данном случае можно применить точный метод II.

Пусть, как и выше, T – число успешных реализаций исследуемого события. Известно, что T распределено по биномиальному закону с параметрами $(n,\ p)$: $\mathbf{P}_p\{T<t\}=\mathrm{Bin}(t\,|\,n,p)$, где p – вероятность этого события при однократном наблюдении. Это распределение убывает с ростом p. Поэтому, если t – экспериментальное значение статистики T, то

а) нижняя (1-α)-доверительная граница $\underline{p}$ для вероятности успеха p есть решение уравнения $Bin(t\,|\,n,\underline{p})=1-\alpha$;

б) верхняя (1-α)-доверительная граница $\overline{p}$ для вероятности успеха p есть решение уравнения $Bin(t+1\,|\,n,\overline{p})=\alpha$.

Для особо любопытных приведем

доказательство корректности метода II.

Пусть $\underline{\theta} = \underline{\theta}(t)$ - решение уравнения $\mathbf{P}_{\underline{\theta}}\{T < t\} = 1-\alpha$, и константа

$$\tilde{t}(\theta) = \sup\{\, t:\ \mathbf{P}_{\theta}\{T < t\} \le 1-\alpha \ \}.$$

В силу непрерывности слева функции распределения, в точке $\tilde{t}$ также выполняется неравенство $\mathbf{P}_{\theta}\{T < \tilde{t}(\theta)\} \le 1-\alpha$. С другой стороны, для непрерывной справа функции $\mathbf{P}_{\theta}\{T \le t\}$ имеет место обратное неравенство

$$\mathbf{P}_{\theta}\{T \le \tilde{t}(\theta)\} \ge 1-\alpha. \qquad (*)$$

Таким образом, справедлива следующая цепочка соотношений:

$$\underline{\theta}(t) \le \theta \quad \Leftrightarrow \quad \mathbf{P}_{\theta}\{T < t\} \le 1-\alpha \quad \Leftrightarrow \quad t \le \tilde{t}(\theta)\ .$$

Первая эквивалентность здесь справедлива в силу монотонного убывания функции распределения по параметру и по определению $\underline{\theta}$, вторая – в силу определения $\tilde{t}$. Соотношение (*) гарантирует теперь, что вероятность первого неравенства $\mathbf{P}_{\theta}\{\,\underline{\theta}(T) \le \theta\} = \mathbf{P}_{\theta}\{T \le \tilde{t}(\theta)\} \ge 1-\alpha$, что и требовалось доказать. Аналогично доказывается утверждение для верхней границы.

Глава VI. Исследование зависимости между двумя характеристиками

Очень часто в эксперименте наблюдается не одна, а несколько характеристик одного и того же объекта (например, рост и вес человека, урожайность зерна и количество внесенных удобрений и т.п.). Предполагается, что значения характеристик изменяются от объекта к объекту случайным образом. Требуется

1) выяснить, имеется ли зависимость между исследуемыми характеристиками;
2) построить уравнение наилучшего прогноза одной характеристики по значениям другой.

Теоретические основы.

Обозначим через X, Y наблюдаемые в эксперименте случайные величины. Попытаемся сначала спрогнозировать возможное значение характеристики Y, если известно наблюдённое значение характеристики $X = x$. Если прогноз осуществляется с помощью некоторой функции $h(x)$, то мерой качества такого прогноза служит среднеквадратическая ошибка $\mathbf{E}(Y - h(X))^2$. Функцию $h^*(x)$, для которой достигается минимум среднеквадратической ошибки, называют функцией *регрессии* Y на X. Теоретический вид этой функции весьма прост:

$h^*(x) = E\{Y \mid X = x\}$ – условное среднее Y при фиксированном значении с.в. $X = x$.

На практике построение хорошей оценки для неё возможно лишь для дискретных с.в. X. Поэтому обычно рассматривают регрессию не среди всех возможных функций, а только среди линейных. Уравнение *линейной*

среднеквадратической *регрессии* Y на X – линейной функции, минимизирующей среднеквадратическую ошибку, имеет вид

$$y = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_x),$$

где μ_X, σ_X^2 – среднее значение и дисперсия с.в. X,

μ_Y, σ_Y^2 – среднее значение и дисперсия с.в. Y,

$\rho = \dfrac{E\{(X-\mu_X)(Y-\mu_Y)\}}{\sigma_X\,\sigma_Y}$ – коэффициент *корреляции* между Y и X.

Коэффициент корреляции ρ часто называют коэффициентом линейной связности. Такая интерпретация обусловлена следующими свойствами этого коэффициента.

1) Он принимает значения от -1 до 1 и не зависит от масштаба измерений.
2) Если $\rho = \pm 1$, то между с.в. Y и X существует точная линейная связь, причем при $\rho = 1$ эта связь имеет положительную направленность (с ростом одной характеристики, растет и другая), а при $\rho = -1$ – отрицательную.
3) Если с.в. Y и X независимы, то $\rho = 0$.
4) Для нормального случайного вектора равенство нулю коэффициента корреляции эквивалентно независимости с.в. Y и X.
5) Минимальная среднеквадратическая ошибка линейного прогноза характеристики Y по значениям характеристики X равна $\sigma_Y^2(1-\rho^2)$.
6) Если $\rho = 0$, то наилучший прогноз с.в. Y – это ее среднее μ_Y, а линия регрессии Y на X представляет собой прямую, параллельную оси OX.

Свойство 5) применяют при описании степени зависимости между случайными величинами. Считается, что разброс с.в. Y на $\rho^2 \cdot 100\%$ обусловлен влиянием на нее с.в. X и на $(1-\rho^2)\cdot 100\%$ внутренними

факторами, присущими самой с.в. Y, или другими неучтенными факторами.

Неизвестные параметры линейной регрессии легко оцениваются своими выборочными аналогами:

$\bar{x}, s_x^2$ – средним значением и дисперсией выборки X,

$\bar{y}, s_Y^2$ – средним значением и дисперсией выборки Y,

$r = \dfrac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y}$ – выборочным коэффициентом корреляции.

Заметим, что последний коэффициент является состоятельной, но смещенной оценкой истинного коэффициента корреляции ρ.

Таким образом, оценку уравнения регрессии Y на X можно записать в виде

$$y = \tilde{y}(x) = \bar{y} + b_{y/x}(x - \bar{x})$$

с коэффициентом регрессии $b_{y/x} = r \cdot s_y / s_x$. Для этой линии достигается минимум суммы расстояний по оси OY между выборочными точками и графиком линии регрессии, когда минимум ищется среди всех линейных функций:

$$\sum_{i=1}^{n}[(y_i - \tilde{y}(x_i)]^2 = \min_{b,c}\sum_{i=1}^{n}[(y_i - (bx_i + c)]^2 .$$

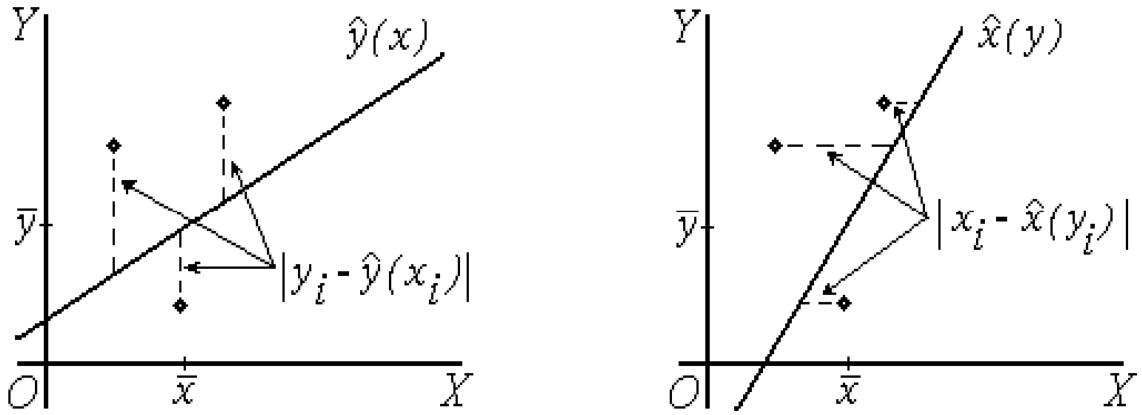
При построении регрессии X на Y можно просто в уравнении регрессии произвести перестановку переменных $x \leftrightarrow y$:

$$x = \tilde{x}(y) = \bar{x} + b_{x/y}(y - \bar{y}), \quad \text{где } b_{x/y} = r \cdot s_x / s_y .$$

Для этой линии достигается минимум суммы расстояний по оси OX между выборочными точками и графиком линии регрессии. Последнее уравнение используется при отыскании наилучшего прогноза характеристики X по наблюдённому значению Y. Для построения графика удобнее привести это уравнение к виду

$$y = \bar{y} + \frac{1}{b_{x/y}}(x - \bar{x}) .$$

Таким образом, обе линии регрессии проходят через точку с координатами $(\bar{x}, \bar{y})$ и отличаются лишь коэффициентом наклона. Можно показать, что в привычной системе координат (x – по оси абсцисс, y – по оси ординат) регрессия X на Y проходит круче регрессии Y на X. Схему построения регрессий Y на X и X на Y иллюстрирует следующий рисунок.



Если выборочный коэффициент корреляции r близок к нулю, то линии регрессии будут близки к взаимно перпендикулярным прямым, проходящим параллельно осям координат, причем регрессия Y на X будет параллельна оси OX. Если r по модулю близок к 1, то угол между линиями регрессии становится близким к нулю, и в пределе (при $|r| = 1$) обе линии совпадут.

На практике чаще всего свойство 4) коэффициента корреляции переносят на все вероятностные модели и интерпретируют равенство нулю коэффициента корреляции как независимость наблюдаемых характеристик, что не всегда верно. В общем случае независимость случайных величин X и Y означает, что

$$\mathbf{P}\{X \in A, Y \in B\} = \mathbf{P}\{X \in A\} \cdot \mathbf{P}\{Y \in B\} \qquad (*)$$

для любых событий A и B из области значений с.в. X и Y. Если известны совместное распределение $F(x, y)$ вектора (X, Y) и частные распределения $F_X(x)$ и $F_Y(y)$ его компонентов, то для проверки независимости достаточно проверить выполнение равенства

$$F(x, y) = F_X(x) F_Y(y).$$

Если случайный вектор (X,Y) имеет не нормальное распределение, то для проверки гипотезы независимости его характеристик необходимо тем или иным способом проверить выполнение соотношения (*).

Задание 14.

Проверить независимость двух характеристик по критерию сопряженности хи-квадрат

Постановка задачи.

По выборке $(x_1, y_1),\ldots,(x_n, y_n)$ из двумерного распределения (не обязательно нормального) проверить гипотезу независимости компонентов наблюдаемого случайного вектора (X, Y).

Теоретические основы.

При отсутствии нормальности распределения вектора (X, Y) для проверки независимости его компонентов применяется критерий сопряженности хи-квадрат. Для построения этого критерия необходимо

1) область значений признака X разбить на r интервалов $A_1^x, A_2^x, ..., A_r^x$, а область значений признака Y на s интервалов $B_1^y, B_2^y, ..., B_s^y$.
2) Для каждого сочетания (i,j) подсчитать количество n_{ij} выборочных данных, для которых, одновременно, признак X попадает в i-ый интервал A_i^x, а признак Y – в m-ый интервал B_m^y. Результаты подсчета свести в *таблицу сопряженности признаков*

X \ Y	1-й	…	s-й	Всего
1-й	n_{11}	…	n_{1s}	$n_{1\bullet}$
…	…	…	…	…
r-й	n_{r1}	…	n_{rs}	$n_{r\bullet}$
Вего	$n_{\bullet 1}$	…	$n_{\bullet s}$	$n_{\bullet\bullet} = n$

где, как обычно, точка $\bullet$ на месте одного из индексов означает сумму всех чисел по этому индексу с фиксированным значением второго индекса. Проще говоря, нужно просуммировать

значения по всем столбцам и строкам таблицы (столбец и строка “Всего”). Число в правой крайней нижней ячейке должно равняться общему объему выборки n.

3) Вычислить статистику критерия сопряженности хи-квадрат

$$X^2 = n\sum_{i=1}^{r}\sum_{m=1}^{s}\frac{\left(\frac{n_{im}}{n} - \frac{n_{i\bullet}}{n}\frac{n_{\bullet m}}{n}\right)^2}{\frac{n_{i\bullet}}{n}\frac{n_{\bullet m}}{n}} \quad \text{èëè} \quad X^2 = \sum_{i=1}^{r}\sum_{m=1}^{s}\frac{\left(n\cdot n_{im} - n_{i\bullet}n_{\bullet m}\right)^2}{n\cdot n_{i\bullet}n_{\bullet m}}.$$

При справедливости гипотезы независимости распределение статистики X^2 может быть аппроксимировано распределением хи-квадрат $K_\nu(x)$ с $\nu = (r-1)(s-1)$ степенями свободы:

$$\mathbf{P}\{X^2 < x\} \approx K_\nu(x) \quad (n \to \infty).$$

Следовательно, если гипотеза независимости отвергается при больших значениях статистики X^2, то при $X^2 = x^2$

4) критический уровень значимости $\alpha_{\text{крит}} \approx 1 - K_\nu(x^2)$.

5) Признаки следует признать независимыми если $\alpha_{\text{крит}} > \alpha$.

Идея критерия сопряженности основана на том, что по закону больших чисел относительная частота

$\frac{n_{im}}{n}$ – есть состоятельная оценка вероятности $\mathbf{P}\{X \in A_i^x, Y \in B_m^y\}$,

а частоты

$\frac{n_{i\bullet}}{n}, \frac{n_{\bullet m}}{n}$ – состоятельные оценки вероятностей $\mathbf{P}\{X \in A_i^x\}$, $\mathbf{P}\{Y \in B_m^y\}$.

Поэтому можно ожидать, что для независимых признаков $\frac{n_{im}}{n} \approx \frac{n_{i\bullet}}{n}\cdot\frac{n_{\bullet m}}{n}$

и поэтому значение статистики X^2 будет “не слишком” большим.

Замечание 1. Как всегда, в критериях хи-квадрат число интервалов разбиения и границы разбиения должны выбираться заранее, до проведения статистического эксперимента. В целях упрощения мы выберем по обоим признакам по 4 интервала

$$(-\infty; z_1], \quad (z_1; z_1 + d], \quad (z_1 + d; z_1 + 2d], \quad (z_1 + 2d; \infty).$$

Значение первой границы z_1 и шаг разбиения d будут даны в задании.

Замечание 2. Критерий “безразличен” к способу получения таблицы сопряженности. Очень часто данные сразу имеют вид такой таблицы. Например, в пособии “Курсовой проект …” [2] рассматривается задача проверки гипотезы независимости уровня образования от количества детей в семье. Данные получены путем обследования некоторой совокупности семей, сгруппированной по двум признакам: уровень образования (две градации, $r = 2$) и число детей (четыре градации, $s = 4$).

Задания 15-16.

Проверить независимость двух характеристик по критерию Стьюдента.

Построить линии регрессии.

Постановка задачи.

По выборке $(x_1, y_1),\ldots,(x_n, y_n)$ из двумерного нормального распределения проверить гипотезу независимости компонентов наблюдаемого случайного вектора (X, Y). Построить линии регрессии одного из признаков по другому признаку. Найти наилучший прогноз признака Y при фиксированном значении признака $X = 120$.

Теоретические основы.

Если вектор (X, Y) имеет нормальное распределение, то независимость его компонентов эквивалентна равенству нулю коэффициента корреляции. Поэтому для проверки гипотезы независимости можно проверить гипотезу $\mathbf{H}_0 : \rho = 0$ о коэффициенте корреляции ρ.

Преобразование Стьюдента для выборочной корреляции

$$T = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2}$$

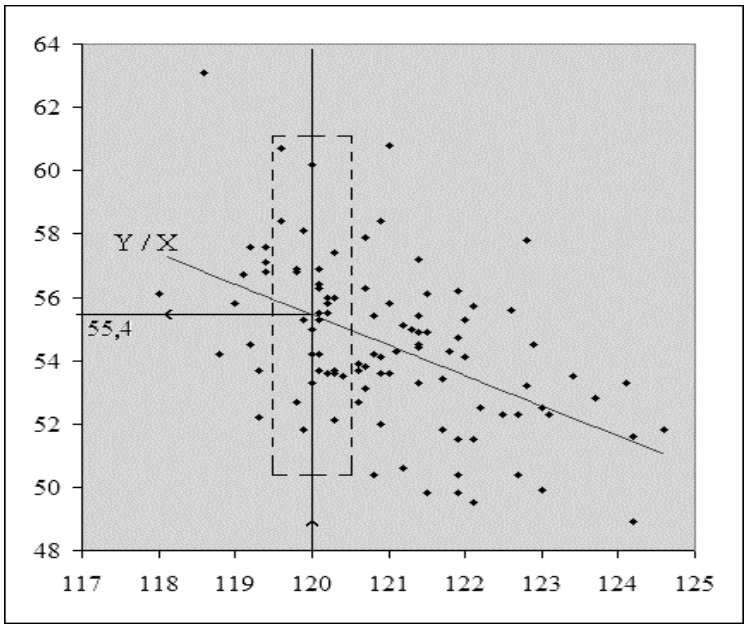
при выборе из двумерного нормального распределения и в условиях гипотезы $\mathbf{H}_0$ имеет распределение Стьюдента $S_{n-2}(t)$ с (n-2)-мя степенями свободы. Таким образом, если нулевую гипотезу отвергать при значениях статистики Стьюдента, в ту или иную сторону (в зависимости от альтернативы) отличающихся от нуля, то критический уровень значимости может быть найден по следующей схеме:

Альтернатива	$\alpha_{\text{крит}}$	пояснение
$\mathbf{H}_1 : \rho \neq 0$	$2(1-S_{n-2}(t))$	$\mathbf{P}\{\mid T \mid > t\}$
$\mathbf{H}_1 : \rho < 0$	$1-S_{n-2}(t)$	$\mathbf{P}\{T > t\}$
$\mathbf{H}_1 : \rho > 0$	$S_{n-2}(t)$	$\mathbf{P}\{T < t\}$.

Замечание 2. Здесь следует подчеркнуть различие между *статистической* и *практической значимостью* коэффициента корреляции. Статистическая значимость коэффициента корреляции означает лишь, что наших данных достаточно для подтверждения зависимости между исследуемыми признаками. Практическая значимость при этом будет означать, что эти признаки могут быть достаточно точно спрогнозированы один по другому. Таким образом, если для практической значимости необходимо, чтобы истинный коэффициент корреляции был очень большим ($\pm 0{,}7$ и выше), то для статистической значимости может оказаться достаточным проведение большого числа наблюдений при очень маленьком коэффициенте корреляции.

Проиллюстрируем эти положения на примере данных из задания 16, обработанных в пособии [4]. В этом пособии при выполнении задания 16 получено значение $r = -0{,}5$ при объеме выборки $n = 101$. Таким образом, критический уровень значимости $\alpha_{\text{крит}} < 0{,}001$, что свидетельствует об очень высокой статистической значимости, однако в этом случае только 25% изменчивости каждого из признаков (см. свойство 5 коэффициента корреляции) можно объяснить влиянием на него другого признака. Другими словами, хотя зависимость между признаками и есть, однако она имеет низкую практическую значимость.

Графически это можно проиллюстрировать следующим образом. Проведем на графике линий регрессии вертикальную линию из точки x=120 (см. рисунок ниже).



Точка пересечения этой линии с прямой регрессии Y на X даст наилучший прогноз значения признака Y при значении признака X=120:

$$Y = -0{,}961*120+170{,}817 = 55{,}4.$$

Реальные значения, близкие к вертикальной линии x=120 (точки на рисунке, обведенные прямоугольным контуром), имеют большой разброс по вертикали, что говорит о плохом качестве прогноза.