**Лекция 13. Корпеляционный анализ. Применение корреляционного анализа в психологии**

Применение статистических методов при обработке материалов психологических исследований дает большую возможность извлечь из экспериментальных данных полезную информацию. Одним из самых распространенных методов статистики является корреляционный анализ.

Термин «корреляция» впервые применил французский палеонтолог Ж. Кювье, который вывел «закон корреляции частей и органов животных» (этот закон позволяет восстанавливать по найденным частям тела облик всего животного). В статистику указанный термин ввел английский биолог и статистик Ф. Гальтон (не просто «связь» – *relation*, а «как бы связь» – *corelation*).

**Корреляционный анализ** – это проверка гипотез о связях между переменными с использованием коэффициентов корреляции, двумерной описательной статистики, количественной меры взаимосвязи (совместной изменчивости) двух переменных. Таким образом, это совокупность методов обнаружения корреляционной зависимости между случайными величинами или признаками.

**Корреляционный анализ для двух случайных величин заключает в себе:**

- построение корреляционного поля и составление корреляционной таблицы;

- вычисление выборочных коэффициентов корреляции и корреляционных отношений;

-проверку статистической гипотезы значимости связи.

Основное назначение корреляционного анализа – выявление связи между двумя или более изучаемыми переменными, которая рассматривается как совместное согласованное изменение двух исследуемых характеристик. Данная изменчивость обладает тремя основными характериcтиками: формой, направлением и силой.

По форме корреляционная связь может быть линейной или нелинейной. Более удобной для выявления и интерпретации корреляционной связи является линейная форма. Для линейной корреляционной связи можно выделить два основных направления: положительное («прямая связь») и отрицательное («обратная связь»).

Сила связи напрямую указывает, насколько ярко проявляется совместная изменчивость изучаемых переменных. В психологии функциональная взаимосвязь явлений эмпирически может быть выявлена только как вероятностная связь соответствующих признаков. Наглядное представление о характере вероятностной связи дает диаграмма рассеивания – график, оси которого соответствуют значениям двух переменных, а каждый испытуемый представляет собой точку.



В качестве числовой характеристики вероятностной связи используют коэффициенты корреляции, значения которых изменяются в диапазоне от –1 до +1. После проведения расчетов исследователь, как правило, отбирает только наиболее сильные корреляции, которые в дальнейшем интерпретируются (табл. 1).



Критерием для отбора «достаточно сильных» корреляций может быть как абсолютное значение самого коэффициента корреляции (от 0,7 до 1), так и относительная величина этого коэффициента, определяемая по уровню статистической значимости (от 0,01 до 0,1), зависящему от размера выборки. В малых выборках для дальнейшей интерпретации корректнее отбирать сильные корреляции на основании уровня статистической значимости. Для исследований, которые проведены на больших выборках, лучше использовать абсолютные значения коэффициентов корреляции.

Таким образом, задача корреляционного анализа сводится к установлению направления (положительное или отрицательное) и формы (линейная, нелинейная) связи между варьирующими признаками, измерению ее тесноты, и, наконец, к проверке уровня значимости полученных коэффициентов корреляции.

В настоящее время разработано множество различных коэффициентов корреляции. Наиболее применяемыми являются *r*-Пирсона, *r*-Спирмена и *τ*-Кендалла. Современные компьютерные статистические программы в меню «Корреляции» предлагают именно эти три коэффициента, а для решения других исследовательских задач предлагаются методы сравнения групп.

Выбор метода вычисления коэффициента корреляции зависит от типа шкалы, к которой относятся переменные (табл. 2).



Для переменных с интервальной и с номинальной шкалой используется коэффициент корреляции Пирсона (корреляция моментов произведений). Если, по меньшей мере, одна из двух переменных имеет порядковую шкалу или не является нормально распределенной, используется ранговая корреляция по Спирмену или

t-Кендалла. Если же одна из двух переменных является дихотомической, можно использовать точечную двухрядную корреляцию (в статистической компьютерной программе SPSS эта возможность отсутствует, вместо нее может быть применен расчет ранговой корреляции). В том случае если обе переменные являются дихотомическими, используется четырехполевая корреляция (данный вид корреляции рассчитываются SPSS на основании определения мер расстояния и мер сходства). Расчет коэффициента корреляции между двумя недихотомическими переменными возможен только тогда, кода связь между ними линейна (однонаправлена). Если связь, к примеру, *U*-образная (неоднозначная), коэффициент корреляции не пригоден для использования в качестве меры силы связи: его значение стремится к нулю.

Таким образом, условия применения коэффициентов корреляции будут следующими:

- переменные, измеренные в количественной (ранговой, метрической) шкале на одной и той же выборке объектов;

- связь между переменными является монотонной.

**Основная статистическая гипотеза,** которая проверяется корреляционным анализом, является ненаправленной и содержит утверждение о равенстве корреляции нулю в генеральной совокупности *H0: rxy* = 0. При ее отклонении принимается альтернативная гипотеза *H1: rxy* ≠ 0 о наличии положительной или отрицательной корреляции – в зависимости от знака вычисленного коэффициента корреляции.

На основании принятия или отклонения гипотез делаются содержательные выводы. Если по результатам статистической проверки *H0: rxy* = 0 не отклоняется на уровне a, то содержательный вывод будет следующим: связь между *X* и *Y* не обнаружена. Если же при *H0 rxy* = 0 отклоняется на уровне a, значит, обнаружена положительная (отрицательная) связь между *X* и *Y*. Однако к интерпретации выявленных корреляционных связей следует подходить осторожно. С научной точки зрения, простое установление связи между двумя переменными не означает существования причинно-следственных отношений. Более того, наличие корреляции не устанавливает отношения последовательности между причиной и следствием. Оно просто указывает, что две переменные взаимосвязаны между собой в большей степени, чем это можно ожидать при случайном совпадении. Тем не менее, при соблюдении осторожности применение корреляционных методов при исследовании причинно-следственных отношений вполне оправдано. Следует избегать категоричных фраз типа «переменная X является причиной увеличения показателя *Y*». Подобные утверждения следует формулировать как предположения, которые должны быть строго обоснованы теоретически.

Подробное описание математической процедуры для каждого коэффициента корреляции дано в учебниках по математической статистике [3]; [4]; [8]; [11] и др. Мы же ограничимся описанием возможности применения этих коэффициентов в зависимости от типа шкалы измерения.

**Корреляция метрических переменных**

Для изучения взаимосвязи двух метрических переменных, измеренных на одной и той же выборке, применяется коэффициент корреляции *r*-Пирсона. Сам коэффициент характеризует наличие только линейной связи между признаками, обозначаемыми, как правило, символами *X* и *Y*. Коэффициент линейной корреляции является параметрическим методом и его корректное применение возможно только в том случае, если результаты измерений представлены в шкале интервалов, а само распределение значений в анализируемых переменных отличается от нормального в незначительной степени. Существует множество ситуаций, в которых его применение целесообразно. Например: установление связи между интеллектом школьника и его успеваемостью; между настроением и успешностью выхода из проблемной ситуации; между уровнем дохода и темпераментом и т. п.

**Коэффициент Пирсона** находит широкое применение в психологии и педагогике. Например, в работах И. Я. Каплуновича [6, с. 115] и П. Д. Рабиновича, М. П. Нуждиной [9, с. 112] для подтверждения выдвинутых гипотез был использован расчет коэффициента линейной корреляции Пирсона.

При обработке данных «вручную» необходимо вычислить коэффициент корреляции, а затем определить *p*-уровень значимости (в целях упрощения проверки данных пользуются таблицами критических значений *rxy*, которые составлены с помощью этого критерия). Величина коэффициента линейной корреляции Пирсона не может превышать +1 и быть меньше чем –1. Эти два числа +1 и –1 являются границами для коэффициента корреляции. Когда при расчете получается величина, большая +1 или меньшая –1, это свидетельствует, что произошла ошибка в вычислениях.

При вычислениях на компьютере статистическая программа (SPSS, Statistica) сопровождает вычисленный коэффициент корреляции более точным значением *p*-уровня.

Для статистического решения о принятии или отклонении *H0* обычно устанавливают *α*= 0,05, а для большого объема наблюдений (100 и более) *α*= 0,01. Если *p ≤ α, H0* отклоняется и делается содержательный вывод, что обнаружена статистически достоверная (значимая) связь между изучаемыми переменными (положительная или отрицательная – в зависимости от знака корреляции). Когда *p > α, H0* не отклоняется, содержательный вывод ограничен констатацией, что связь (статистически достоверная) не обнаружена.

Если связь не обнаружена, но есть основания полагать, что связь на самом деле есть, следует проверить возможные причины недостоверности связи.

**Нелинейность связи** – для этого проанализировать график двумерного рассеивания. Если связь нелинейная, но монотонная, перейти к ранговым корреляциям. Если связь не монотонная, то делить выборку на части, в которых связь монотонная, и вычислить корреляции отдельно для каждой части выборки, или делить выборку на контрастные группы и далее сравнивать их по уровню выраженности признака.

**Наличие выбросов и выраженная асимметрия распределения одного или обоих признаков.** Для этого необходимо посмотреть гистограммы распределения частот обоих признаков. При наличии выбросов или асимметрии исключить выбросы или перейти к ранговым корреляциям.

**Неоднородность выборки** (проанализировать график двумерного рассеивания). Попытаться разделить выборку на части, в которых связь может иметь разные направления.

Если же связь статистически достоверна, то прежде чем делать содержательный вывод, необходимо исключить возможность ложной корреляции:

*- связь обусловлена выбросами*. При наличии выбросов перейти к ранговым корреляциям или исключить выбросы;

*- связь обусловлена влиянием третьей переменной*. Если есть подобное явление, необходимо вычислить корреляцию не только для всей выборки, но и для каждой группы в отдельности. Если «третья» переменная метрическая – вычислить частную корреляцию.

**Коэффициент частной корреляции *rxy-z*** вычисляется в том случае, если необходимо проверить предположение, что связь между двумя переменными *X* и *Y* не зависит от влияния третьей переменной *Z*. Очень часто две переменные коррелируют друг с другом только за счет того, что обе они согласованно меняются под влиянием третьей переменной. Иными словами, на самом деле связь между соответствующими свойствами отсутствует, но проявляется в статистической взаимосвязи под влиянием общей причины. Например, общей причиной изменчивости двух переменных может являться возраст при изучении взаимосвязи различных психологических особенностей в разновозрастной группе. При интерпретации частной корреляции с позиции причинности следует быть осторожным, так как если *Z* коррелирует и с *X* и с *Y*, а частная корреляция *rxy-z* близка к нулю, из этого не обязательно следует, что именно *Z* является общей причиной для *X* и *Y*.

Корреляция ранговых переменных

Если к количественным данным неприемлем коэффициент корреляции *r*-Пирсона, то для проверки гипотезы о связи двух переменных после предварительного ранжирования могут быть применены корреляции *r*-Спирмена или *τ*-Кендалла. Например, в исследовании психофизических особенностей музыкально одаренных подростков И. А. Лавочкина [7, с. 149] был использован критерий Спирмена.

Для корректного вычисления обоих коэффициентов (Спирмена и Кендалла) результаты измерений должны быть представлены в шкале рангов или интервалов. Принципиальных отличий между этими критериями не существует, но принято считать, что коэффициент Кендалла является более «содержательным», так как он более полно и детально анализирует связи между переменными, перебирая все возможные соответствия между парами значений. Коэффициент Спирмена более точно учитывает именно количественную степень связи между переменными.

**Коэффициент ранговой корреляции Спирмена** является непараметрическим аналогом классического коэффициента корреляции Пирсона, но при его расчете учитываются не связанные с распределением показатели сравниваемых переменных (среднее арифметическое и дисперсия), а ранги. Например, необходимо определить связь между ранговыми оценками качеств личности, входящими в представление человека о своем «Я реальном» и «Я идеальном».

**Коэффициент Спирмена** широко используется в психологических исследованиях. Например, в работе Ю. В. Бушова и Н. Н. Несмеловой [1]: для изучения зависимости точности оценки и воспроизведения длительности звуковых сигналов от индивидуальных особенностей человека был использован именно он.

Так как этот коэффициент – аналог *r*-Пирсона, то и применение его для проверки гипотез аналогично применению коэффициента *r*-Пирсона. То есть проверяемая статистическая гипотеза, порядок принятия статистического решения и формулировка содержательного вывода – те же. В компьютерных программах (SPSS, Statistica) уровни значимости для одинаковых коэффициентов *r*-Пирсона и *r*-Спирмена всегда совпадают.

Преимущество коэффициента *r*-Спирмена по сравнению с коэффициентом *r*-Пирсона – в большей чувствительности к связи. Мы используем его в следующих случаях:

- наличие существенного отклонения распределения хотя бы одной переменной от нормального вида (асимметрия, выбросы);

- появление криволинейной (монотонной) связи.

Ограничением для применения коэффициента *r*-Спирмена являются:

- по каждой переменной не менее 5 наблюдений;

- коэффициент при большом количестве одинаковых рангов по одной или обеим переменным дает огрубленное значение.

**Коэффициент ранговой корреляции *τ*-Кендалла** является самостоятельным оригинальным методом, опирающимся на вычисление соотношения пар значений двух выборок, имеющих одинаковые или отличающиеся тенденции (возрастание или убывание значений). Этот коэффициент называют еще *коэффициентом конкордации*. Таким образом, основной идеей данного метода является то, что о направлении связи можно судить, попарно сравнивая между собой испытуемых: если у пары испытуемых изменение по *X* совпадает по направлению с изменением по *Y*, это свидетельствует о положительной связи, если не совпадает – об отрицательной связи, например, при исследовании личностных качеств, имеющих определяющее значение для семейного благополучия. В этом методе одна переменная представляется в виде монотонной последовательности (например, данные мужа) в порядке возрастания величин; другой переменной (например, данные жены) присваиваются соответствующие ранговые места. Количество инверсий (нарушений монотонности по сравнению с первым рядом) используется в формуле для корреляционных коэффициентов.

При подсчете *τ-*Кендалла «вручную» данные сначала упорядочиваются по переменной *X*. Затем для каждого испытуемого подсчитывается, сколько раз его ранг по *Y*оказывается меньше, чем ранг испытуемых, находящихся ниже. Результат записывается в столбец «Совпадения». Сумма всех значений столбца «Совпадение» и есть *P*– общее число совпадений, подставляется в формулу для вычисления коэффициента Кендалла, который более прост в вычислительном отношении, но при возрастании выборки, в отличие от*r*-Спирмена, объем вычислений возрастает не пропорционально, а в геометрической прогрессии. Так, например, при *N* = 12 необходимо перебрать 66 пар испытуемых, а при *N* = 489 – уже 1128 пар, т. е. объем вычислений возрастает более чем в 17 раз. При вычислениях на компьютере в статистической программе (SPSS, Statistica) коэффициент Кендалла обсчитывается аналогично коэффициентам *r*-Спирмена и *r*-Пирсона. Вычисленный коэффициент корреляции *τ*-Кендалла характеризуется более точным значением *p*-уровня.

**Применение коэффициента Кендалла** является предпочтительным, если в исходных данных имеются выбросы.

Особенностью ранговых коэффициентов корреляции является то, что максимальным по модулю ранговым корреляциям (+1, –1) не обязательно соответствуют строгие прямо или обратно пропорциональные связи между исходными переменными *X* и *Y*: достаточна лишь монотонная функциональная связь между ними. Ранговые корреляции достигают своего максимального по модулю значения, если большему значению одной переменной всегда соответствует большее значение другой переменной (+1), или большему значению одной переменной всегда соответствует меньшее значение другой переменной и наоборот (–1).

Проверяемая статистическая гипотеза, порядок принятия статистического решения и формулировка содержательного вывода те же, что и для случая *r*-Спирмена или *r*-Пирсона.

Если статистически достоверная связь не обнаружена, но есть основания полагать, что связь на самом деле есть, следует сначала перейти от коэффициента

*r*-Спирмена к коэффициенту *τ*-Кендалла (или наоборот), а затем проверить возможные причины недостоверности связи:

*нелинейность связи*: для этого посмотреть график двумерного рассеивания. Если связь не монотонная, то делить выборку на части, в которых связь монотонная, или делить выборку на контрастные группы и далее сравнивать их по уровню выраженности признака;

*неоднородность выборки*: посмотреть график двумерного рассеивания, попытаться разделить выборку на части, в которых связь может иметь разные направления.

Если же связь статистически достоверна, то прежде чем делать содержательный вывод, необходимо исключить возможность ложной корреляции (по аналогии с метрическими коэффициентами корреляции).

**Корреляция дихотомических переменных**

При сравнении двух переменных, измеренных в дихотомической шкале, мерой корреляционной связи служит так называемый коэффициент j, который представляет собой коэффициент корреляции для дихотомических данных.

Величина коэффициента φ лежит в интервале между +1 и –1. Он может быть как положительным, так и отрицательным, характеризуя направление связи двух дихотомически измеренных признаков. Однако интерпретация φ может выдвигать специфические проблемы. Дихотомические данные, входящие в схему вычисления коэффициента φ, не похожи на двумерную нормальную поверхность, следовательно, неправильно считать, что интерпретируемые значения *rxy*=0,60 и φ = 0,60 одинаковы. Коэффициент φ можно вычислить методом кодирования, а также используя так называемую четырехпольную таблицу или таблицу сопряженности.

Для применения коэффициента корреляции φ необходимо соблюдать следующие условия:

- сравниваемые признаки должны быть измерены в дихотомической шкале;

- число варьирующих признаков в сравниваемых переменных *X* и *Y* должно быть одинаковым.

Данный вид корреляции рассчитывают в компьютерной программе SPSS на основании определения мер расстояния и мер сходства. Некоторые статистические процедуры, такие как факторный анализ, кластерный анализ, многомерное масштабирование, построены на применении этих мер, а иногда сами представляют добавочные возможности для вычисления мер подобия.

В тех случаях когда одна переменная измеряется в дихотомической шкале (переменная *X*), а другая в шкале интервалов или отношений (переменная *Y*), используется бисериальный коэффициент корреляции, например, при проверке гипотез о влиянии пола ребенка на показатель роста и веса. Этот коэффициент изменяется в диапазоне от –1 до +1, но его знак для интерпретации результатов не имеет значения. Для его применения необходимо соблюдать следующие условия:

- сравниваемые признаки должны быть измерены в разных шкалах: одна *X* – в дихотомической шкале; другая *Y* – в шкале интервалов или отношений;

- переменная *Y* имеет нормальный закон распределения;

- число варьирующих признаков в сравниваемых переменных *X*и *Y* должно быть одинаковым.

Если же переменная *X* измерена в дихотомической шкале, а переменная *Y* в ранговой шкале (переменная *Y*), можно использовать рангово-бисериальный коэффициент корреляции, который тесно связан с τ-Кендалла и использует в своем определении понятия совпадения и инверсии. Интерпретация результатов та же.

**Проведение корреляционного анализа с помощью компьютерных программ SPSS и Statistica** – простая и удобная операция. Для этого после вызова диалогового окна Bivariate Correlations (Analyze>Correlate> Bivariate…) необходимо переместить исследуемые переменные в поле Variables и выбрать метод, с помощью которого будет выявляться корреляционная связь между переменными. В файле вывода результатов для каждого рассчитываемого критерия содержится квадратная таблица (Correlations). В каждой ячейке таблицы приведены: само значение коэффициента корреляции (Correlation Coefficient), статистическая значимость рассчитанного коэффициента Sig, количество испытуемых.

В шапке и боковой графе полученной корреляционной таблицы содержатся названия переменных. Диагональ (левый верхний – правый нижний угол) таблицы состоит из единиц, так как корреляция любой переменной с самой собой является максимальной. Таблица симметрична относительно этой диагонали. Если в программе установлен флажок «Отмечать значимые корреляции», то в итоговой корреляционной таблице будут отмечены статистически значимые коэффициенты: на уровне 0,05 и меньше – одной звездочкой (\*), а на уровне 0,01 – двумя звездочками (\*\*).

Итак, подведем итоги: основное назначение корреляционного анализа – это выявление связи между переменными. Мерой связи являются коэффициенты корреляции, выбор которых напрямую зависит от типа шкалы, в которой измерены переменные, числа варьирующих признаков в сравниваемых переменных и распределения переменных. Наличие корреляции двух переменных еще не означает, что между ними существует причинная связь. Хотя корреляция прямо не указывает на причинную связь, она может быть ключом к разгадке причин. На ее основе можно сформировать гипотезы. В некоторых случаях отсутствие корреляции имеет более глубокое воздействие на гипотезу о причинной связи. Нулевая корреляция двух переменных может свидетельствовать, что никакого влияния одной переменной на другую не существует.

**Литература**

1. Бушов Ю. В., Несмелова Н. Н. Зависимость точности оценки и воспроизведения длительности звуковых сигналов от индивидуальных особенностей человека // Вопросы психологии. 1996. № 3.

2. Бююль А., Цёфель П. SPSS: Искусство обработки информации. М., 2002.

3. Глас Дж., Стенли Дж. Статистические методы в педагогике и психологии. М., 1976.

4. Ермолаев О. Ю. Математическая статистика для психологов. М., 2003.

5. Калинин С. И. Компьютерная обработка данных для психологов. СПб., 2004.

6. Каплунович И. Я. Содержание мыслительных операций в структуре пространственного мышления // Вопросы психологии. 1987. № 6.

7. Лавочкина И. А. Психофизические особенности музыкально одаренных подростков // Вопросы психологии. 1988. № 4.

8. Наследов А. Д. Математические методы психологического исследования. Анализ и интерпретация данных. СПб., 2004.

9. Рабинович П. Д., Нуждина М. П. Зависимость успеваемости студентов от их характерологических особенностей // Вопросы психологии. 1987. № 6.

10. Романко В. К. Курс теории вероятностей и математической статистики для психологов. М., 2000.

11. Сидоренко Е. В. Методы математической обработки в психологии. СПб., 2001.

12. Солсо Р. Л., Джонсон Х. Х., Бил М. К. Экспериментальная психология: практический курс. СПб., 2001.