ФЕДЕРАЛЬНОЕ АГЕНТСТВО НАУЧНЫХ ОРГАНИЗАЦИЙ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ УЧРЕЖДЕНИЕ НАУКИ
БОТАНИЧЕСКИЙ САД-ИНСТИТУТ
УФИМСКОГО НАУЧНОГО ЦЕНТРА РОССИЙСКОЙ АКАДЕМИИ НАУК

Реферативная работа по дисциплине
«Биометрия»
на тему:

«Корреляционный анализ»

Выполнила: аспирантка 1 курса
очного обучения Давлетбаева С.Ф.
Проверил: к.б.н., с.н.с. лаборатории
дикорастущей флоры и интродукции
травянистых растений Жигунов О.Ю.

5 (отл.)

Уфа-2015

СОДЕРЖАНИЕ:

ВВЕДЕНИЕ **3**

1.Корреляция **4**

1.1.Понятие корреляции **4**

1.2.Свойства корреляционной связи **5**

1.3. Коэффициенты ковариации и корреляции **6**

2. Параметрические показатели корреляции **8**

2.1.Коэффициент линейной корреляции Пирсона **8**

3.Непараметрические показатели **10**

3.1.Коэффициент ранговой корреляции Спирмена **10**

3.2.Коэффициент ранговой корреляции Кендалла **11**

3.3. Коэффициент корреляции знаков Фехнера **12**

Заключение **13**

Литература **14**

ВВЕДЕНИЕ

Статическая обработка цифрового материала, полученного путем опытов, учетов и наблюдений при биологических исследованиях, необходима для проверки степени достоверности результатов и правильного их обобщения. Особенно важна роль статистических методов, как средства, помогающего принять верное решение в условиях неопределенности. Корреляционный анализ является одним из наиболее широко используемых статистических методов.

Целью корреляционного анализа является измерение статистической взаимозависимости между двумя или более переменными.

Между величинами может существовать более точная, или функциональная, связь, когда одному значению аргумента (x) соответствует одно определенное значение функции (y), и менее точная – корреляционная связь, когда одному конкретному значению аргумента соответствует приближенное значение, или некоторое множество значений функции, в той или иной степени близких друг к другу. Близость в этом множестве значений функции друг к другу соответствует понятию силы, или тесноты корреляционной связи. Чем больше сила корреляционной связи, тем ближе эта связь к функциональной.

Объекты ботанических исследований (популяции, особи, органы, клетки) всегда в той или иной степени неоднородны по своим наследственным особенностям и, кроме того, в любом, даже тщательно поставленном эксперименте его объекты испытывают неучитываемые воздействия многих факторов внешней среды. Поэтому между признаками объектов ботанических исследований бывают исключительно корреляционные связи.

1.Корреляция

1.1.Понятие корреляции

Величины, характеризующие различные свойства объектов, могут быть независимыми или взаимосвязанными. Различают два вида зависимостей между величинами (факторами): функциональную и статистическую.

При функциональной зависимости двух величин значению одной из них обязательно соответствует одно или несколько точно определенных значений другой величины. Функциональная связь двух факторов возможна лишь при условии, что вторая величина зависит только от первой и не зависит ни от каких других величин. Функциональная связь одной величины с множеством других возможна, если эта величина зависит только от этого множества факторов. В реальных ситуациях существует бесконечно большое количество свойств самого объекта и внешней среды, влияющих друг на друга, поэтому такого рода связи не существуют, иначе говоря, функциональные связи являются математическими абстракциями. Их применение допустимо тогда, когда соответствующая величина в основном зависит от соответствующих факторов.

При исследовании функционирования систем многие параметры следует считать случайными, что исключает проявление однозначного соответствия значений. Воздействие общих факторов, наличие объективных закономерностей в поведении объектов приводят лишь к проявлению статистической зависимости. Статистической называют зависимость, при которой изменение одной из величин влечет изменение распределения других (другой), и эти другие величины принимают некоторые значения с определенными вероятностями. Функциональную зависимость в таком случае следует считать частным случаем статистической: значению одного фактора соответствуют значения других факторов с вероятностью, равной единице (Задорина, 2009).

Корреляция зависимость - статистическая взаимосвязь двух или более случайных величин (либо величин, которые можно с некоторой

допустимой степенью точности считать таковыми). При этом изменения значений одной или нескольких из этих величин сопутствуют систематическому изменению значений другой или других величин (Шмойлова, 2002).

Если же у взаимосвязанных величин вариацию имеет только одна переменная, а другая является детерминированной, то такую связь называют не корреляционной, а регрессионной (Гусаров, 2001).

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» – *co-relation*) в конце) XIX века (Задорина, 2009; Елисеева, 2002) Однако точную формулу для подсчета коэффициента корреляции разработал его ученик Карл Пирсон.

Наличие корреляционной зависимости между переменными не всегда означает наличие непосредственной связи этих величин друг с другом: наблюдаемая связь часто существует благодаря другим переменным (не двум рассматриваемым), а изучаемые величины могут быть связаны между собой через латентные (скрытые от исследователя) переменные.

Корреляционная связь, в отличие от функциональной, показывает лишь тенденцию изменения одной величины под действием другой, следовательно, на основании корреляции можно утверждать лишь о степени связи между переменными, но не о существовании причинно- следственной зависимости между ними.

1.2.Свойства корреляционной связи

Корреляционные связи различаются по тесноте (силе) связи и количеству признаков. По тесноте (силе) корреляционной связи принято выделять: 1) функциональную, 2) тесную (сильную), 3) среднюю (умеренную), 4) слабую и 5) нулевую (отсутствующую) виды связи. По

количеству признаков корреляция может быть парной (между двумя признаками) и множественной (между несколькими признаками). Форма парной корреляции может быть линейной, описываемой линейной функцией регрессии, и нелинейной (криволинейной), описываемой нелинейными функциями. Парная линейная корреляция, в свою очередь, может быть положительной («прямой») и отрицательной («обратной»). При положительной корреляции при возрастании одного признака в среднем увеличивается другой, в случае же отрицательной корреляции при возрастании одного признака другой в среднем уменьшается (Харченко, 2008).

1.3. Коэффициенты ковариации и корреляции

Для оценки тесноты и направления связи между изучаемыми переменными при их стохастической зависимости пользуются показателями ковариации и корреляции.

Ковариацией $cov(X,Y)$ случайных величин X и Y называется среднее произведение отклонений каждой пары значений величин X и Y в исследуемых массивах данных:

$$cov(X,Y) = \overline{(X_i - \overline{X})(Y_i - \overline{Y})} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}).$$

Ковариация есть характеристика системы случайных величин, описывающая помимо рассеивания величин X и Y еще и линейную связь между ними.

Доказано, что для независимых случайных величин X и Y их ковариация равна нулю, а для зависимых случайных величин она отличается от нуля (хотя и необязательно). Поэтому ненулевое значение ковариации означает зависимость случайных величин. Однако обращение в нуль ковариации не гарантирует независимости, бывают зависимые случайные величины, ковариация которых равна нулю.

Из формулы определения ковариации видно, что ковариация характеризует не только зависимость величин, но и их рассеивание. Действительно, если, например, одна из величин *X* или *Y* мало отличается от своего математического ожидания (почти не случайна), то показатель ковариации будет мал, какой бы тесной зависимостью ни были связаны величины *X* и *Y*. Так что обращение в нуль ковариации величин *X* и *Y* является недостаточным условием для их независимости, а только необходимым.

Использование ковариации в качестве меры связи признаков не совсем удобно, так как показатель ковариации не нормирован и при переходе к другим единицам измерения меняет значение (Ходасевич, 2000).

Нормированное значение коэффициента ковариации называется коэффициентом корреляции:

$$\mathbf{r}_{XY} = \frac{\mathbf{cov}_{XY}}{\sigma_X \sigma_Y} = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum(X-\bar{X})^2 \sum(Y-\bar{Y})^2}}.$$

где σx, σy – средние квадратические отклонения переменных X и Y; X, Y – их математические ожидания. Коэффициент корреляции представляет собой безразмерную величину, изменяющуюся в пределах от –1 до 1. Значение коэффициента корреляции выражает лишь долю от максимально возможной ковариации, в чем и состоит его преимущество перед коэффициентом ковариации. Величина коэффициента корреляции не меняется при увеличении или уменьшении на одно и то же число или в одно и то же число раз всех значений переменных.

При ρ = ±1 корреляционная связь представляет собой линейную функциональную зависимость, при этом все наблюдаемые значения располагаются на прямой. При ρ = 0 линейная корреляционная связь отсутствует, корреляционное поле представляет собой круг. Чем ближе значение | ρ | к единице, тем с большим основанием можно считать, что

изучаемые величины находятся в линейной зависимости. Для независимых случайных величин коэффициент корреляции равен нулю. Однако равенство нулю коэффициента корреляции не всегда означает независимость случайных величин: оно свидетельствует лишь об отсутствии линейной корреляционной зависимости между изучаемыми величинами, но не корреляционной зависимости вообще. Коэффициент корреляции может быть равен нулю в случае нелинейной связи.

Направление линейной корреляционной связи определяется зна- ком коэффициента корреляции: для «прямой», положительной связи $\rho > 0$, для «обратной», отрицательной связи $\rho < 0$. Теснота (сила) линейной связи между случайными величинами оп- ределяется абсолютной величиной коэффициента корреляции:

$|\rho| = 1$ – величины связаны линейной функциональной зависимостью;

$0{,}95 \le |\rho| < 1$ – связь очень сильная, практически функциональная;

$0{,}75 \le |\rho| < 0{,}95$ – связь тесная (сильная);

$0{,}5 \le |\rho| < 0{,}75$ – связь средняя (умеренная);

$0{,}2 \le |\rho| < 0{,}5$ – связь слабая;

$0 \le |\rho| < 0{,}2$ – практически нет связи.

Коэффициент корреляции связан с корреляционным отношением следующим образом: $0 \le |\rho| \le \eta \le 1$. В случае линейной связи они равны: $|\rho| = \eta$, поэтому: 1) расхождение между коэффициентом корреляции и корреляционным отношением используется в качестве критерия линейности корреляционной зависимости; 2) в случае линейной связи коэффициент детерминации равен квадрату коэффициента корреляции (Харченко, 2008).

2.Параметрические показателя корреляции

2.1.Коэффициент линейной корреляции Пирсона

Коэффициент корреляции Пирсона (r-Пирсона) - мера корреляции, подходящая для двух непрерывных (метрических переменных), измеренных на одной и той же выборке. Он позволяет определить, насколько

пропорциональная изменчивость двух переменных (Наследов, 2004). Данный коэффициент разработали Карл Пирсон, ФрэнсисЭджуорт и Рафаэль Уэлдон в 90-х годах XIX века. Коэффициент корреляции изменяется в пределах отминус единицы до плюс единицы (Гмурман, 2004).

Линейный коэффициент корреляции Пирсона (rp) используется для измерения тесноты связи между двумя количественными признаками X и Y. Расчет коэффициента может производиться только при соблюдении следующих условий:

- обе переменные являются количественными и непрерывными.

- как минимум один из признаков (а лучше оба) имеет нормальное распределение.

- зависимость между переменными носит линейный характер.

- гомоскедастичность (вариабельность одной переменной не зависит от значений другой переменной).

- независимость участников исследования друг от друга.

- парность наблюдений (признак X и признак Y изучаются у одних и тех же участников исследования) (Унгуряну, 2014).

Формула расчет коэффициента корреляции Пирсона следующая:

$$r_{xy} = \frac{\Sigma(x_i - \bar{x}) \times (y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2 \times \Sigma(y_i - \bar{y})^2}}$$

где x_i — значения переменной X; y_i — значения переменной Y; $\bar{x}$ — среднее арифметическое для переменной X; $\bar{y}$ -среднее арифметическое для переменной Y.

Величина коэффициента корреляции варьируется от 0 до 1.Величина коэффициента корреляции отражает силы связи. При оценке силы связи коэффициентов корреляции используется шкала Чеддока: от 0 до 0,3 - очень слабая; от 0,3 до 0,5 – слабая; от 0, 5 до 0,7 – средняя; от 0,7 до 0, 9 – высокая; от 0,9 до 1 - очень высокая.

При отрицательной корреляции значения силы связи между переменными меняют на противоположные. *Например:* если величина коэффициента корреляции между переменными -0,36, то это слабая

отрицательная корреляция, и скорее всего мы не будем принимать ее в расчет;если величина коэффициента корреляции равна 0, следовательно переменные не связаны между собой; если величина коэффициента корреляции между переменными равна 0,25 то это очень слабая корреляция и в большинстве случаев мы не берем ее в расчет; если величина коэффициента корреляции между переменными равна 0, 75 то это высокая корреляция и в своих интерпретациях нам стоит обратить на нее внимание; если величина коэффициента корреляции равна 1, следовательно полностью взаимосвязаны.

3.Непараметрические показатели

3.1.Коэффициент ранговой корреляции Спирмена

Коэффициент корреляции рангов, предложенный К. Спирменом, относится к непараметрическим показателям связи между переменными, измеренными в ранговой шкале. При расчете этого коэффициента не требуется никаких предположений о характере распределений признаков в генеральной совокупности. Этот коэффициент определяет степень тесноты связи порядковых признаков, которые в этом случае представляют собой ранги сравниваемых величин.

При использовании коэффициента ранговой корреляции условно оценивают тесноту связи между признаками, считая значения коэффициента равные 0,3 и менее, показателями слабой тесноты связи; значения более 0,4, но менее 0,7 - показателями умеренной тесноты связи, а значения 0,7 и более - показателями высокой тесноты связи (Лезин, 2014).

В принципе число ранжируемых признаков может быть любым, но сам процесс ранжирования большего, чем 20 числа признаков - затруднителен. Возможно, что именно поэтому таблица критических значений рангового коэффициента корреляции рассчитана лишь для сорока ранжируемых признаков ($n < 40$).

Ранговый коэффициент корреляции Спирмена подсчитывается по формуле:

$$\rho = 1 - \frac{6 \cdot \sum d^2}{n \cdot (n^2 - 1)}$$

где n - количество ранжируемых признаков (показателей, испытуемых); D - разность между рангами по двум переменным для каждого испытуемого; $\sum d^2$- квадратов разностей рангов (Кобзарь).

3.2. Коэффициент ранговой корреляции Кендалла

Коэффициент корреляции Кенделла - мера линейной связи между случайными величинами. Корреляция Кенделла является ранговой, то есть для оценки силы связи используются не численные значения, а соответствующие им ранги. Коэффициент инвариантен по отношению к любому монотонному преобразованию шкалы измерения.

Заданы две выборки $x = (x_1, \ldots, x_n),\ y = (y_1, \ldots, y_n)$.

Коэффициент корреляции Кенделла вычисляется по формуле:

$\tau = 1 - \frac{4}{n(n-1)} R$, где $R = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left[[x_i < x_j] \neq [y_i < y_j] \right]$

количество инверсий, образованных величинами y_i, расположенными в порядке возрастания соответствующих x_i.

Коэффициент τ принимает значения из отрезка $[-1;\ 1]$. Равенство $\tau = 1$ указывает на строгую прямую линейную зависимость, $\tau = -1$ на обратную.Обоснование критерия Кенделла:пары $(x_i,\ y_i)$ и $(x_j,\ y_j)$ согласованы,если $x_i < x_j$ и $y_i < y_j$ или $x_i > x_j$ и $y_i > y_j$,то есть $sign(x_j - x_i) sign(y_j - y_i) = 1$

Пусть S - число согласованных пар, R - число несогласованных пар. Тогда, в предположении, что среди x_i и среди y_i нет совпадений, превышение согласованности над несогласованностью есть:

$$T = S - R = \sum_{i<j} sign(x_j - x_i) sign(y_j - y_i)$$

Для измерения степени согласия Кенделл предложил следующий коэффициент:

$$\tau = \frac{T}{maxT} = \frac{2T}{n(n-1)} = \frac{2(S-R)}{n(n-1)} = 1 - \frac{4}{n(n-1)} R.$$

Таким образом, коэффициент τ (линейно связанный с R) можно считать *мерой неупорядоченности* второй последовательности относительно первой (Кобзарь, 2006).

3.3.Коэффициент корреляции знаков Фехнера

Коэффициент Фехнера, характеризует элементарную степень тесноты связи, которую целесообразно использовать для установления факта наличия связи, когда существует небольшой объем исходной информации.Он основан на сравнении поведения отклонений индивидуальных значений каждого признака (x и y) от своей средней величины. При этом во внимание принимаются не величины отклонений $(x_i - \bar{x})$ и $(y_i - \bar{y})$, а их знаки («+» или «-»). Определив знаки отклонения от средней величины в каждом ряду, рассматривают все пары знаков и подсчитывают число их совпадений ($\sum C$) и несовпадений ($\sum H$).Коэффициент Фехнера (K_Φ)рассчитывается как отношение разности чисел пар совпадений и несовпадений знаков к их сумме, т.е. к общему числу наблюдаемых единиц: $K_\Phi = \frac{\sum C - \sum H}{\sum C + \sum H}$. Очевидно, что если знаки всех отклонений по каждому признаку совпадают, то $\sum H = 0$ и тогда $K_\Phi = 1$. Это характеризует наличие прямой связи. Если все знаки не совпадают, то $\sum C = 0$, а $K_\Phi = -1$, что характеризует обратную связь. Коэффициент Фехнера, как и любой другой показатель тесноты связи, может принимать значения от -1 до +1.

Заключение:

Термин «корреляция» впервые применил французский палеонтолог Ж. Кювье. В статистику указанный термин ввел английский биолог и статистик Ф. Гальтон в конце XIX века. Однако точную формулу для подсчета коэффициента корреляции разработал его ученик Карл Пирсон.

Корреляция зависимость - статистическая взаимосвязь двух или более случайных величин (либо величин, которые можно с некоторой допустимой степенью точности считать таковыми). При этом изменения значений одной или нескольких из этих величин сопутствуют систематическому изменению значений другой или других величин. Суть корреляционного анализа заключается в расчете коэффициентов корреляции. Коэффициенты корреляции могут принимать, как правило, положительные и отрицательные значения. Знак коэффициента корреляции позволяет интерпретировать направление связи, а абсолютное значение – силу связи.

Результаты корреляционного анализа могут дать существенную информацию об исследуемом объекте, а также подсказать и направление дальнейших исследований, и совокупность требуемых методов, в том числе статистических, необходимых для более полного изучения объекта.

Популярность метода обусловлена двумя моментами: коэффициенты корреляции относительно просты в подсчете, их применение не требует специальной математической подготовки. В сочетании с простотой интерпретации, простота применения коэффициента привела к его широкому распространению в сфере анализа статистических данных.

Корреляционный анализ используется как составная часть более сложных, комплексных методов, в том числе факторного анализа, некоторых версий кластер-анализа и другие.

Литература:

1. Гмурман В. Е. Теория вероятностей и математическая статистика: Учебное пособие для вузов. - 10-е издание, стереотипное. - Москва: Высшая школа, 2004. - 479 с.
2. Гусаров В.М. «Теория статистики», М.: ЮНИТИ, 2001. – 247 с.
3. Елисеева И. И., Юзбашев М. М. Общая теория статистики: Учебник - 4-е издание, переработанное и дополненное. - Москва: Финансы и Статистика, 2002. - 480 с.
4. Задорина Н.А. Обработка экспериментальных данных на ЭВМ.- Рыбинск: РГАТА, 2009. - 101с.
5. Кобзарь А. И. Прикладная математическая статистика. — М.: Физматлит, 2006. - 624-626 с.
6. Лезин И.А., Маркелов Д.Е. Извлечение знаний из больших массивов данных / Известия Самарского научного центра Российской академии наук, Т. 16, №4(2), 2014.
7. Наследов, А. Д. Математические методы психологического исследования. Анализ и интерпретация данных. – СПб.: Речь, 2004. – 392 с.
8. Суслов В. И., Ибрагимов Н. М., Талышева Л. П., Цыплаков А. А. Эконометрия. - Новосибирск: СО РАН, 2005. - 744 с.
9. Унгуряну Т. Н., Гржибовский А. М. Корреляционный анализ с использованием пакета статистических программ STATA // Экология человека, 2014. № 9. С. 60–64.
10. Ходасевич Г.Б. Обработка экспериментальных данных на ЭВМ. Учебное пособие. — СПб.: Санкт-Петербургский государственный университет телекоммуникаций им. проф. М.А.Бонч-Бруевича, 2000. — 100 с.
11. Шмойлова Р. А. Общая теория статистики: Учебник 3-е издание, переработанное. - Москва: Финансы и Статистика, 2002. - 560 с.
12. Харченко М.А. Корреляционный анализ: Учебное пособие для вузов. - Воронеж: Изд-во ВГУ, 2008. - 31 с.