Содержание

Введение

Исследователя нередко интересует, как связаны между собой две или большее количество переменных в одной или нескольких изучаемых выборках. Например, может ли рост влиять на вес человека или может ли давление влиять на качество продукции?

Такого рода зависимость между переменными величинами называется корреляционной, или корреляцией. Корреляционная связь - это согласованное изменение двух признаков, отражающее тот факт, что изменчивость одного признака находится в соответствии с изменчивостью другого.

Известно, например, что в среднем между ростом людей и их весом наблюдается положительная связь, и такая, что чем больше рост, тем больше вес человека. Однако из этого правила имеются исключения, когда относительно низкие люди имеют избыточный вес, и, наоборот, астеники, при высоком росте имеют малый вес. Причиной подобных исключений является то, что каждый биологический, физиологический или психологический признак определяется воздействием многих факторов: средовых, генетических, социальных, экологических и т.д.

Корреляционные связи - это вероятностные изменения, которые можно изучать только на представительных выборках методами математической статистики. Оба термина - корреляционная связь и корреляционная зависимость - часто используются как синонимы. Зависимость подразумевает влияние, связь - любые согласованные изменения, которые могут объясняться сотнями причин. Корреляционные связи не могут рассматриваться как свидетельство причинно-следственной зависимости, они свидетельствуют лишь о том, что изменениям одного признака, как правило, сопутствуют определенные изменения другого.

Корреляционная зависимость - это изменения, которые вносят значения одного признака в вероятность появления разных значений другого признака.

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.

Объектом исследования в данной курсовой работе является процесс хранения рыбы.

В роли же предмета исследования выступает изменение сроков хранения рыбы в зависимости от температуры хранения.

Цель моей курсовой работы: проведение корреляционного анализа зависимости сроков хранения рыбы от изменения температуры хранения.

Исходя из поставленной цели, формируются следующие задачи:

1. Оценить математические ожидания, дисперсии, средние квадратические отклонения и коэффициент корреляции случайных величин X и Y.
2. Построить (аналитически и графически) регрессионные модели $Y = \hat{y}(x) + \varepsilon$. Найти доверительный интервал для условного математического ожидания. Оценить корреляционное отношение для параболической регрессии.
3. Дать интервальную оценку случайной величины Y. Определить толерантный интервал.

Для достижения вышеперечисленных задач в курсовой работе были использованы следующие методы: оценивание параметров, интервальное оценивание, метод построения однофакторных регрессионных моделей, метод наименьших квадратов.

Постановка практической задачи

При мониторинге условий хранения рыбы, была выявлена зависимость, представленная в Таблице 1.

Изменение сроков хранения рыбы фиксируется в месяцах.

Отклонение от t воздуха (градусы)	-10	-6	-3	-1	0	1	3	6
Изменение сроков хранения рыбы	-24	-12	-4	0	2	4	7	12

Увеличение сроков хранения на данную продукцию происходит с уменьшением температуры воздуха. При увеличении температуры в морозильной камере на определенное количество градусов, происходит уменьшение срока хранения рыбы на соответствующее количество месяцев. Таким образом, при выполнении вычитания из срока хранения соответствующей цифры и получения отрицательного числа стоит говорить о том, что рыба уже испортилась.

Требуется:

1. Оценить математические ожидания, дисперсии, средние квадратические отклонения и коэффициент корреляции случайных величин X, Y. На координатной плоскость нанести точки из таблицы.

2. Используя методы корреляционного анализа построить (аналитически и графически) регрессионные модели $Y=\hat{y}(x)+\varepsilon$, причём в качестве эмпирического уравнения регрессии $\hat{y}(x)$ взять линейную функцию и параболу. Для линейного случая прямым методом построить два уравнения регрессии (Y на X, X на Y), причём среднее квадратическое отклонение случайной величины ε оценить двумя способами. Найти доверительный интервал для условного математического ожидания $M[Y/x]$ с доверительной

вероятностью 1-α=0,95 при предположении о нормальном условном распределении случайной величины Y. Для параболической регрессии оценить корреляционное отношение.

3. Дать интервальную оценку случайной величины Y с вероятностью попадания в интервал p=0,96, если взятое из той же генеральной совокупности значение x_{n+1}=-8, при предположении, что эмпирическое уравнение регрессии построено точно. Определить толерантный интервал.

Применение методов корреляционного анализа к решению задачи

Данные в виде выборки из генеральной совокупности представлены в Таблице 2.

Таблица 2. Выборка из генеральной совокупности

x_i	-10	-6	-3	-1	0	1	3	6
y_i	-24	-12	-4	0	2	4	7	12

1.Оценить математические ожидания, дисперсии, средние квадратические отклонения и коэффициент корреляции случайных величин X, Y.

1.1 Оценки математических ожиданий:

$$\widehat{M}X \equiv \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\widehat{\mathrm{M}}X = \frac{1}{8}(-10 + (-6) + (-3) + (-1) + 0 + 1 + 3 + 6) = -1{,}25$$

$$\widehat{M}Y \equiv \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$\widehat{\mathrm{M}}Y = \frac{1}{8}(-24 + (-12) + (-4) + 0 + 2 + 4 + 7 + 12) = -1{,}875$$

1.2 Несмещённые оценки дисперсий:

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

$$\hat{S}_x^2 = \frac{1}{7}((-10 + 1.25)^2 + (-6 + 1.25)^2 + (-3 + 1.25)^2 + (-1 + 1.25)^2 + (0 + 1.25)^2 + (1 + 1.25)^2 + (3 + 1.25)^2 + (6 + 1.25)^2) = 25.64$$

$$\hat{S}_y^2 = \frac{1}{7}((-24 + 1.875)^2 + (-12 + 1.875)^2 + (-4 + 1.875)^2 + (0 + 1.875)^2 + (2 + 1.875)^2 + (4 + 1.875)^2 + (7 + 1.875)^2 + (12 + 1.875)^2) = 131.55$$

1.3 Оценки средних квадратических отклонений (С.К.О):

$$\hat{S}_X = \sqrt{\hat{S}_X^2}$$

$$\hat{S}_x = \sqrt{25.64} = 5.06$$

$$\hat{S}_y = \sqrt{131.55} = 11.47$$

1.4 Оценка ковариации.

$$\hat{\text{cov}}(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

$$\hat{\text{cov}}(X,Y) = \frac{1}{7}\big((-10+1.25)(-24+1.88) + (-6+1.25)(-12+1.88) + (-3+1.25)(-4+1.88) + (-1+1.25)(0+1.88) + (0+1.25)(2+1.88) + (1+1.25)(4+1.88) + (3+1.25)(7+1.88) + (6+1.25)(12+1.88)\big) = 57.46$$

1.5Оценка коэффициента корреляции.

$$\hat{\rho} = \frac{\hat{\text{cov}}(X,Y)}{\hat{S}_X \cdot \hat{S}_Y}, \hat{\rho} = \frac{57.46}{5.06*11.47} = 0.99$$

Так как оценка коэффициента корреляции $|\hat{\rho}| > 0{,}9$, сила связи между x_i и y_i высокая. Гипотеза H_0: ρ =0 не подтвердилась. А значит, между ***X*** и ***Y*** существует линейная статистическая связь.

Чем ближе модуль коэффициента корреляции к 1, тем сильнее связь между ***X*** и ***Y***.

1.6 Графическое представление данных

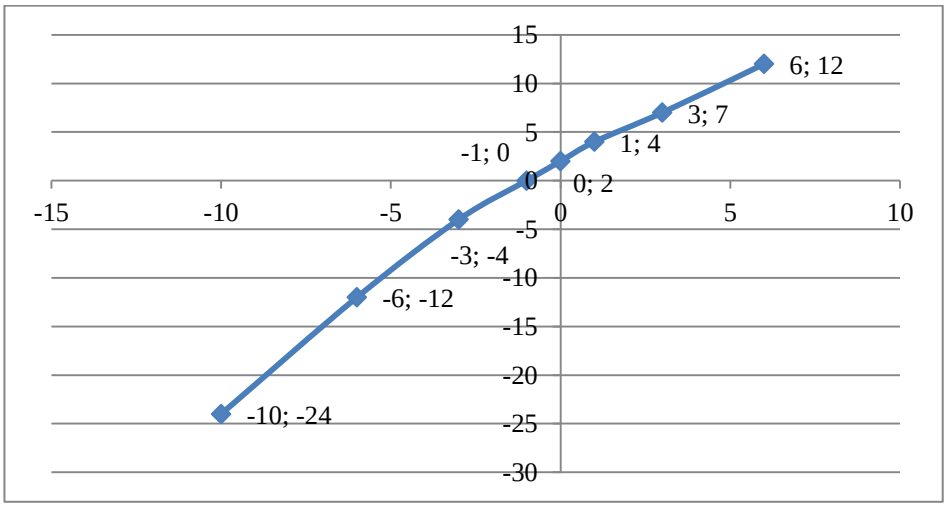

Рис.1 Исходные данные из таблицы

2.Регрессионные модели.

Построим регрессионную модель: $Y = \hat{y}(x) + \varepsilon$, где

$\hat{y}$(x) – оценка условного математического ожидания случайной величины Y (при фиксированном значении x)

ε – случайная величина, распределённая по нормальному закону с $M\varepsilon=0$ и дисперсией $D\varepsilon$.

2.1 Линейная функция.

Для построения первой регрессионной модели возьмем линейную функцию:

$$\hat{y}(x) = ax + b$$

Для построения используем прямой метод.

Построим уравнение X на Y по формуле:

$$\hat{x}(y) = \bar{x} + \hat{\rho} \cdot \frac{\hat{s}_x}{\hat{s}_y} \cdot (y - \bar{y})$$

Для вычисления используем несмещенные оценки средних квадратических отклонений и коэффициентов корреляции.

$$\hat{x}(y) = 1{,}25 + 0{,}99 \cdot \frac{5.06}{11.47} \cdot (\text{y} - 1.875) = 0.44\text{y} - 0{,}43$$

Построим уравнение Y на X по формуле:

$$\hat{y}(x) = \bar{y} + \hat{\rho}\frac{\hat{\sigma}_y}{\hat{\sigma}_x}(x_i - \bar{x})$$

$$\hat{y}\ (\text{x}) = 1.875 + 0{,}99 \cdot \frac{11.47}{5.06} \cdot \left(x_i - 1.25\right) = 2{,}24\text{x} - 0.93$$

Построим графики полученных уравнений.

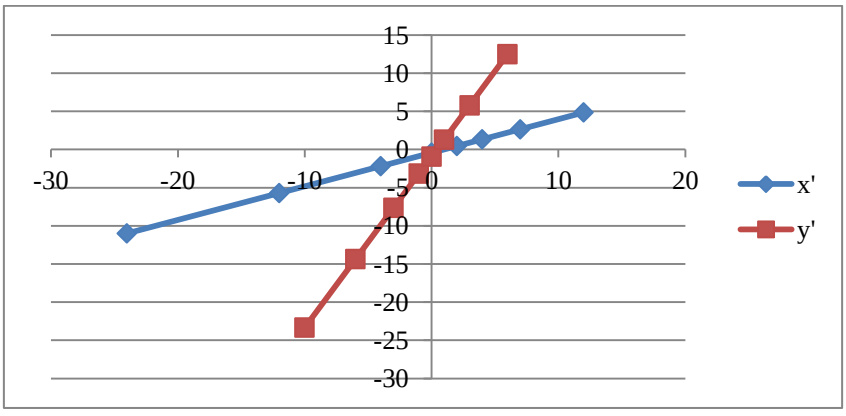

Рис.2 Графики функций регрессии X на Y и Y на X.

Для проверки правильности построения обеих моделей приведём отдельные графики с подписями наносимых значений.

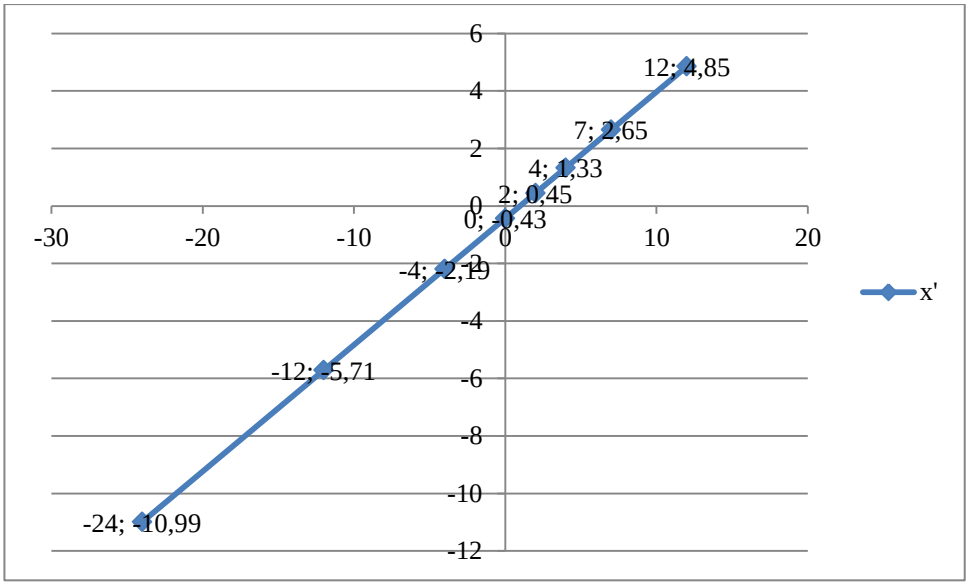

Рис.3 График X на Y

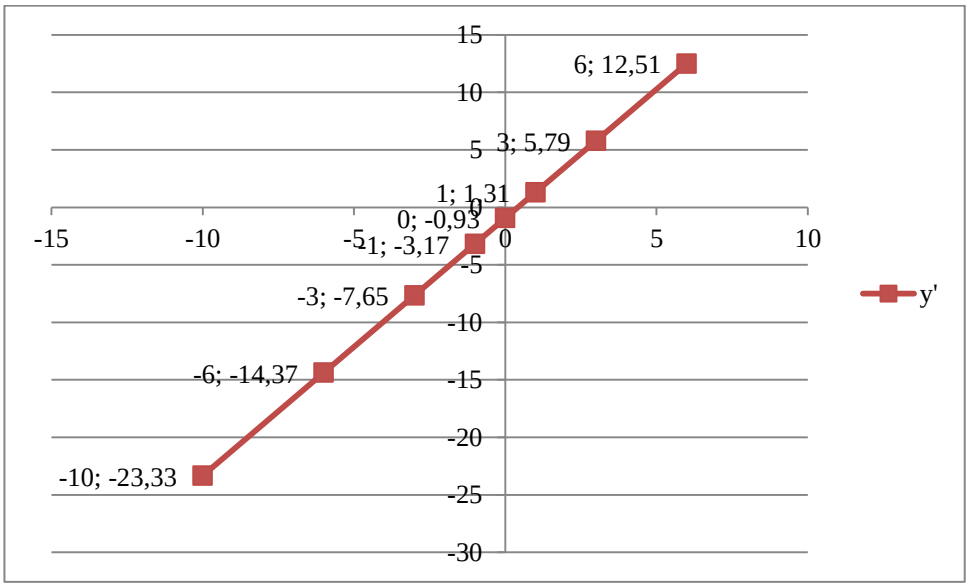

Рис.4 График Y на X.

Теперь найдём оценку среднего квадратического отклонения ($\hat{\sigma}$) случайной величины ε.

1) Первый способ:

$$\hat{\sigma} = \hat{S}_y \sqrt{(1-\hat{\rho}^2)\frac{n-1}{n-2}}$$

Подставим известные значения:

$$\hat{\sigma}=11.47\sqrt{(1-0.99^2)\cdot\frac{7}{6}}=1.23$$

2) Второй способ (для данного случая l=2):

$$\hat{\sigma}=\sqrt{\frac{1}{n-l}\cdot\sum_{i=1}^{n}(y_i-\hat{y}(x_i))^2}$$

Подставим значения:

$$\hat{\sigma}=$$

$$=\sqrt{\frac{1}{8-2}\cdot((-24+23{,}33)^2+(-12+14{,}37)^2+(-4+7{,}65)^2+(0+3{,}17)^2+(2+0{,}93)^2+}$$

$$\sqrt{+(4-1{,}31)^2+(7-5{,}79)^2+(12-12{,}51)^2)}=1.23$$

$\hat{\sigma}$ =1.23

Результаты, полученные с помощью первого и второго методов, равны, значит оценка среднего квадратического отклонения вычислена верно.

2.2 Параболическая функция:

В качестве эмпирической функции возьмём уравнение параболы: $y_{cp}(x)=ax^2+bx+c$

$$Q(\hat{a},\hat{b},\hat{c})=\sum(y_i-\hat{a}x_i^2-\hat{b}x_i-c)^2\to\min_{\hat{a},\hat{b},\hat{c}}$$

Оценочные значения коэффициентов a, b и c найдём из следующей системы уравнений:

$$\begin{cases}\sum_{i=1}^{n}y_i\cdot x_i^2=\hat{a}\cdot\sum_{i=1}^{n}x_i^4+\hat{b}\cdot\sum_{i=1}^{n}x_i^3+\hat{c}\sum_{i=1}^{n}x_i^2\\ \sum_{i=1}^{n}y_i\cdot x_i=\hat{a}\cdot\sum_{i=1}^{n}x_i^3+\hat{b}\cdot\sum_{i=1}^{n}x_i^2+\hat{c}\sum_{i=1}^{n}x_i\\ \sum_{i=1}^{n}y_i=\hat{a}\cdot\sum_{i=1}^{n}x_i^2+\hat{b}\cdot\sum_{i=1}^{n}x_i+\hat{c}\cdot n\end{cases}$$

$$\begin{cases} -15 = \hat{a} \cdot 8 + \hat{b} \cdot (-10) + \hat{c} \cdot 192 \\ 421 = \hat{a} \cdot (-10) + \hat{b} \cdot 192 + \hat{c} \cdot (-1000) \\ -2369 = \hat{a} \cdot 192 + \hat{b} \cdot (-1000) + \hat{c} \cdot 12756 \end{cases}$$

Решим систему линейных уравнений методом Крамера.

Выпишем главный определитель:

$$\Delta = \begin{matrix} 8 & -10 & 192 \\ -10 & 192 & -1000 \\ 192 & -1000 & 12756 \end{matrix} = 7079728$$

Запишем определитель для вычисления $\hat{a}$:

$$\Delta_1 = \begin{matrix} -15 & -10 & 192 \\ 421 & 192 & -1000 \\ -2369 & -1000 & 12756 \end{matrix} = -54024408$$

Запишем определитель для вычисления $\hat{b}$:

$$\Delta_2 = \begin{matrix} 8 & -15 & 192 \\ -10 & 421 & -1000 \\ 192 & -2369 & 12756 \end{matrix} = 14005544$$

Запишем определитель для вычисления $\hat{c}$:

$$\Delta_3 = \begin{matrix} 8 & -10 & -15 \\ -10 & 192 & 421 \\ 192 & -1000 & -2369 \end{matrix} = -439244$$

$$\hat{a} = \frac{\Delta_1}{\Delta} = -0.062$$

$$\hat{b} = \frac{\Delta_2}{\Delta} = 1.978$$

$$\hat{c} = \frac{\Delta_3}{\Delta} = 2.087$$

Уравнение параболы : $\hat{y}(x) = -0.062\ x^2 + 1.978\ x + 2{,}087$

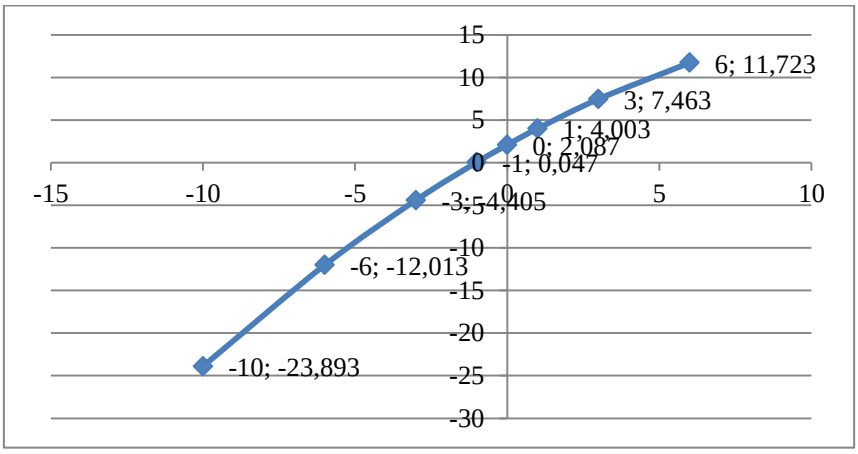

Рис.5 График параболической регрессионной модели

Эмпирическое корреляционное отношение вычисляется для всех форм связи и служит для измерение тесноты зависимости. Изменяется в пределах [0;1].

$$\eta = \sqrt{\frac{\sum(\bar{y}-y_x)^2}{\sum(y_i-\bar{y})^2}} = \sqrt{\frac{920.399}{920.875}} = 0.99948$$

Где $\sum\left(\overline{y}-y_x\right)^2 = 920.875 - 0.476 = 920.399$

При предположении о нелинейной регрессии, величина $\hat{\eta}$ оказалась больше $|\hat{\eta}|$, значит, в качестве $\hat{y}(x)$ используется нелинейная модель.

3. Нахождение доверительного интервала для условного математического ожидания с доверительной вероятностью 1-α=0,95 (нормальное условное распределение случайной величины Y) для регрессионной модели.

$$y_{cp}(x) \in \left[\hat{y}(x) - \Delta_n(x); \hat{y}(x) + \Delta_n(x)\right];$$

$$\Delta_n(x) = \mathrm{t}_\alpha \cdot \frac{\hat{\sigma}}{\sqrt{n}} \cdot \sqrt{1 + \frac{(x-\bar{x})^2}{\hat{S}_x^2 \cdot \frac{n-1}{n}}}$$

Для линейной модели:

По таблице «Критические точки распределения Стьюдента» нельзя найти точное значение для t с доверительной вероятностью 1-α=0,95 при степени свободы k=6, где t – квантиль распределения Стьюдента.

$$t_{0,05} = 2.447$$

$$\Delta = t_{kp} S \sqrt{1 + \frac{1}{n} + \frac{(\bar{x}-x_i)^2}{\sum(x_i-\bar{x})^2}}$$

$t_{крит}$(n-m-1;α/2) = (6;0.025) = 2.447

Расчеты представим в виде таблицы 3.

Таблица 3.

x_i	$y = 0.93 + 2.24x_i$	ε_i	$y_{min} = y - \varepsilon_i$	$y_{max} = y + \varepsilon_i$
-10	-21.48	5.49	-26.97	-15.99
-6	-12.52	4.93	-17.45	-7.59
-3	-5.8	4.71	-10.51	-1.09
-1	-1.31	4.67	-5.99	3.36
0	0.93	4.69	-3.77	5.62
1	3.17	4.73	-1.56	7.9
3	7.65	4.88	2.77	12.53

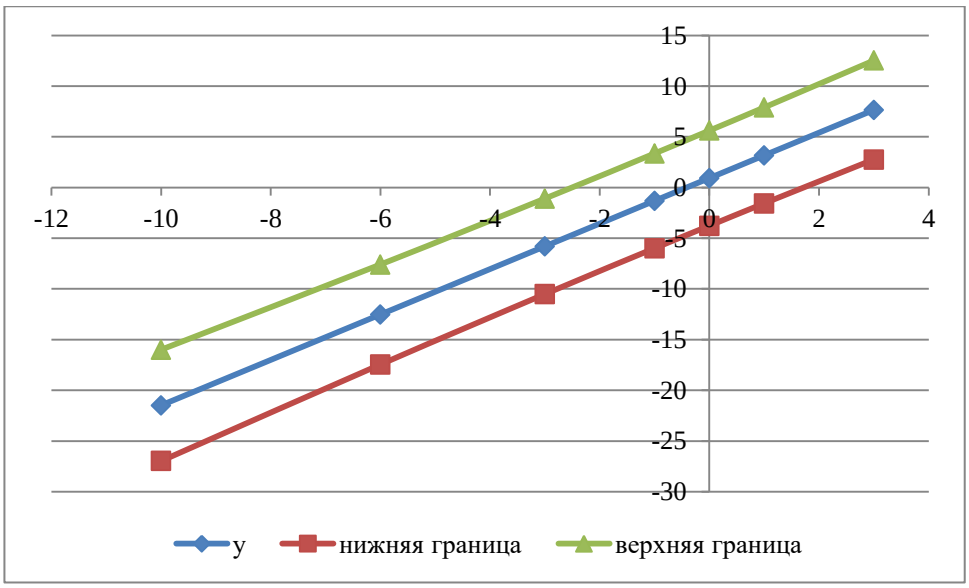

Рис.6 Определение доверительного интервала для линейной регрессионной модели

Определим, в каком интервале будет находиться срок хранения, если отклонение от нормальной температуры составит:

x_{n+1} = x9=-8.

Подставим значение x_{n+1}=-8 в ранее использованную формулу:

y(-8) = 2.241*(-8) + 0.926 = -17.001

Вычислим ошибку прогноза для уравнения y = bx + a

$$\epsilon = t_{kpum} S \sqrt{\frac{1}{n} + \frac{(\overline{x} - x_p)^2}{\sum (x_i - \overline{x})^2}}$$

$$\epsilon = 2.447 \cdot 1.801 \sqrt{\frac{1}{8} + \frac{(-1.25 - (-8))^2}{179.5}} = 2.712$$

-17.001 ± 2.712

(-19.71;-14.29)

С вероятностью 95% можно гарантировать, что значения Y при неограниченно большом числе наблюдений не выйдет за пределы найденных интервалов.

Для нелинейной модели получим:

Стандартная ошибка уравнения.

$$S_y = \sqrt{\frac{\sum(y_i - y_t^2)}{n-m-1}}$$

где m = 2 - количество влияющих факторов в модели тренда.

$$S_y = \sqrt{\frac{0.48}{5}} = 0.31$$

По таблице Стьюдента находим Ттабл

Ттабл (n-m-1;α/2) = (5;0.025) = 2.571

Рассчитаем границы интервала, в котором будет сосредоточено 95% возможных значений Y при неограниченно большом числе наблюдений и t=8

y(tp) ± ε

где

$$\epsilon = t_{kpum} S_y \sqrt{\frac{1}{n} + \frac{(\bar{t} - t_p)^2}{\sum(t_i - \bar{t})^2}}$$

$$\epsilon = 2.571 \cdot 0.309 \sqrt{\frac{1}{8} + \frac{(0-8)^2}{179.5}} = 0.551$$

y(8) = 1.978 -0.062*(-8) + 2.08684514*$(-8)^2$= 136,029

(136,029 - 0.551; 136,029 + 0.551)

(135,478;136,58)

Заключение

Из проведенного корреляционного анализа, можно сделать следующие выводы: изменение срока хранения рыбы и изменение температуры хранения тесно взаимосвязаны.

В ходе интервального оценивания было выяснено, что повышение температуры хранения, приведет к увеличению сроков хранения рыбы.

На основе выборки из генеральной совокупности были произведены расчеты для выявления взаимосвязи между отклонением сроков хранения и изменением температурного режима. Расчеты были произведены с помощью методов корреляционного и регрессионного анализов.

На начальных этапах были найдены оценки математического ожидания, дисперсии, среднего квадратичного отклонения и коэффициента корреляции. Результаты вычислений показали наличие сильной связи.

Затем были построены графики эмпирических уравнений регрессии для линейной и параболической моделей. Также был найден толерантный интервал для p=0,96 для линейной и параболической моделей. ???

Таким образом, на основе анализа исходных данных можно утверждать, что при температуре x_9=-8 и вероятности p=0,96 ??? срок хранения рыбы увеличится и его значение будет лежать в интервале от 20 до 14 месяцев.

Литература

1. Адлер Ю.П., Грановский Ю.В., Маркова Е.В. Планирование эксперимента при поиске оптимальных условий. М.: Наука, 2006.–278 с.
2. Андерсон Т., Введение в многомерный статистический анализ//www.ami.nstu.ru, 2013, 24 с.
3. Бондарь А.Г., Статюха Г.А. Планирование эксперимента в химической технологии. Киев: Высшая школа, 2006 – 335 с.
4. Гмурман В.Е. Руководство к решению задач по теории вероятностей и математической статистике: учеб.пособие. – 11-е изд., перераб. - М.: Высшее образование, 2008. – 404с. – (Основы наук).
5. Ковалев В.В, Волкова О.Н., Анализ хозяйственной деятельности предприятия//polbu.ru, 2010, 2 с.
6. Лекция на тему: "Корреляционный анализ"// www.kgafk.ru, 2006, 8 с.
7. Поляков Л.Е., Коэффициент ранговой корреляции Спирмена//www.eduhmao.ru, 2007, 2 с.