1. Данные в экономике. Классификация данных.

2. Генеральная совокупность и выборка. Суть выборочного метода.

Генеральная совокупность Ω – это совокупность всех подлежащих изучению объектов или явлений. В некоторых задачах генеральную совокупность рассматривают как случайную величину X.
Выборочная совокупность (или выборка) Ω˜ - совокупность случайно отобранных объектов из генеральной совокупности.

- Число объектов совокупности называется объемом совокупности
- Объем генеральной совокупности – N
- Объем выборки – n

Основной метод математической статистики – выборочный. Его суть:

- Выборочный метод – метод матстатистики, где на основе изучения выборки делается заключение обо всей генеральной совокупности.
- Теоретической основой применения выборочного метода является Закон Больших Чисел: при неограниченном увеличении объема выборки её характеристики сколь угодно близко приближаются к характеристикам генеральной совокупности.

3. Способы осуществления выборки. Условия репрезентативности выборки.

Чтобы выборка правильно представляла изучаемый признак генеральной совокупности, хорошо отражала пропорции генеральной совокупности, она должна быть репрезентативной (представительной). Выборка будет репрезентативной, если:

- Её осуществить случайно
- Все объекты генеральной совокупности имеют равные вероятности быть отобранными

Способы формирования выборки:

- Повторный (возвратный): объект после исследования возвращается в генеральную совокупность
- Бесповторный (безвозвратный): объект после исследования не возвращается в генеральную совокупность

4. Понятие вариационного ряда. Дискретные и интервальные статистические ряды: понятие, способы задания.

Статистический ряд – это ранжированный перечень вариантов xi и соответствующих им весов (частот или частостей)
Общий вид статистического ряда частот/частостей

Значение признака xi	x1	x2	…	xk	
Частота ni	n1	n2	…	nk	n1+n2+…+nk=n
Относительная частота ωi	ω1	ω2	…	ωk	ω1+ ω2+…ωk= 1

Где k – число различных вариантов в ряду
Статистические ряды бывают дискретными и интервальными:

- Стат ряд называют дискретным, если любые его варианты отличаются друг от друга на постоянную величину. В таких рядах задаются точечные значения признака.

- Статистический ряд называется интервальным, если любые его варианты отличаются друг от друга на сколь угодно малую величину. Значения признака в них задаются в виде интервалов.

Если число значений признака X велико, то варианты разбивают на отдельные интервалы, т.е. проводят их группировку.

На практике обычно считают, что правильно составленный ряд распределения содержит от 5 до 15 частичных интервалов.

Рекомендуемое число интервалов вычисляется по формуле Стерджеса:

m=1+3,322*lg(n)

Ширина (величина) интервала h равна:

$$h=\frac{x_{max}-x_{min}}{m}$$

5. Эмпирическая функция распределения, её график и свойства.

Эмпирической (статистической) функцией распределения Fn(x) называется функция, равная относительной частоте того, что признак (СВ X) примет значение меньшее заданного действительного числа x, т.е. функция, определяющая, для каждого значения x частость события {X<x}:

$$F_n(x)=\omega(X<x)=\frac{n_i}{n}=\frac{число\ вариант\ x_i\ меньших\ x}{объем\ выборки}$$

x-любое действительное число

Свойства эмпирической функции распределения:

- Значение $F_n(x)$ принадлежат отрезку [0;1]
- $F_n(x)$ является неубывающей функцией
- $F_n(x)=0$ при $x\le xmin$
- $F_n(x)=1$ при $x>xmax$

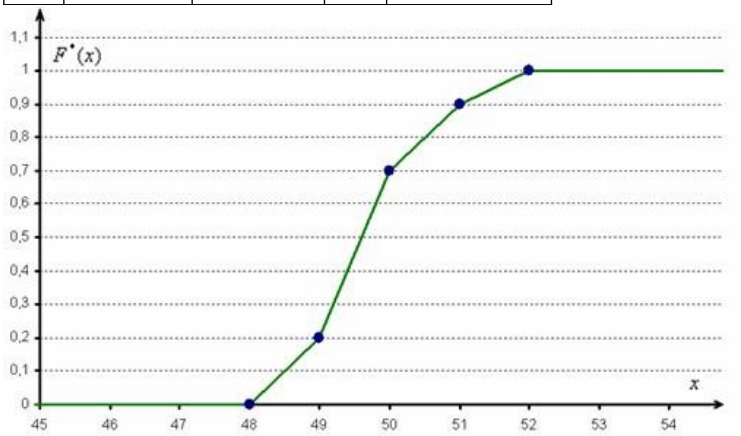
Для дискретного статистического ряда:

xi	x1	x2	…	xk
ωi	ω1	ω2	…	ωk

$$F_n(x)\begin{cases}0\ при\ x\le x1\\ \omega 1\ при\ x1<x\le x2\\ \omega 1+\omega 2\ при\ x2<x\le x3\\ \omega 1+\omega 2+\omega 3\ при\ x3<x\le x4\\ \dots\\ 1\ при\ x>xk\end{cases}$$

Для интервального статистического ряда:

xi	(x1;x2]	(x2;x3]	…	(xk-1;xk]
ω i	ω1	ω2	…	ωk

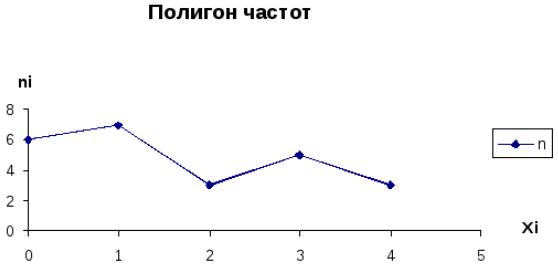


- F(x)=0 при x≤x1
- F(x2)= ω1
- F(x3)= ω1+ ω2
- F(x4)= ω1+ ω2+ ω3
- …
- F(xk)= ω1+ ω2+ …+ ωk-1+ ωk=1
- F(x)=1 при x>xk

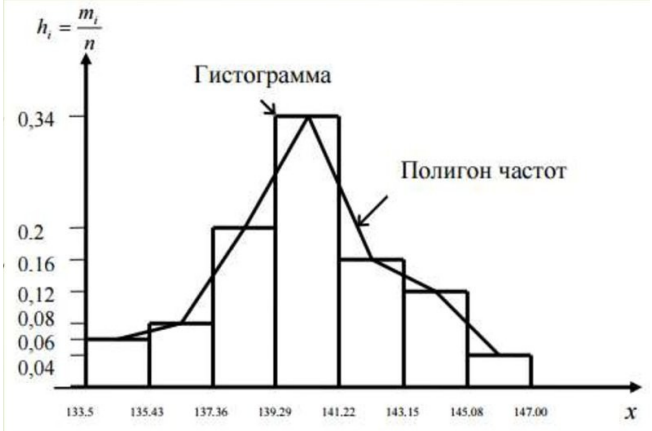
6. Графическое представление статистических рядов: полигон и гистограмма.

Полигон как правило служит для изображения дискретного статистического ряда. Полигон частот (или частостей) – это ломаная, отрезки которой соединяют точки с координатами (xi;ni) или (xi;ωi), i=1,2,…k.

Варианты (xi) откладывают на оси абсцисс, а частоты или частости – на оси ординат.



Гистограмма (т.е. столбчатая диаграмма) служит только для изображения интервальных статистических рядов. Гистограмма частот или частостей – это ступенчатая фигура, состоящая из прямоугольников, основаниями которых служат интервалы (xi; xi+1), i=1,2,…m, а высоты равны частотам (ni) или частостям (ωi).



7. Выбросы. Диаграмма размаха (ящик с усами).

8. Числовые характеристики выборочных распределений: выборочная средняя, мода, медиана, показатели вариации, показатели формы.

Числовые характеристики признака X, рассчитанные по выборке, называются выборочными характеристиками этого признака. Выборочные характеристики являются случайными величинами, а не константами.

Пусть дано стат распределение выборки объема n

А) Дискретный статистический ряд частот:

Значение признака xi	x1	x2	…	xk
Частота ni	n1	n2	…	nk

Б) Интервальный статистический ряд частот

(xi-1;xi)	(x1;x2]	(x2;x3]	…	(xk-1;xk]
ni	n1	n2	…	nk

I Характеристика центра распределения (средние)

- Выборочная средняя – это среднее арифметическое всех значений выборки:
  - Простая – используется, когда данные наблюдения не сведены в вариационный ряд, либо когда все частоты равны 1 или одинаковы

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

  - Взвешенная – используется, когда частоты отличны друг от друга:

$$\overline{x_{выб}} = \frac{\sum_{i=1}^{n} x_i * n_i}{\sum_{i=1}^{k} n_i}$$

- Мода Mo вариационного (статистического) ряда – это вариант, которому соответствует наибольшая частота.
  - Для дискретного вариационного ряда мода равна значению варианты с наибольшей частотой
  - Мода интервального ряда определяется по формуле:

$$Mo = x_{Mo} + \frac{h * n_{Mo} - n_{Mo-1}}{(n¿¿Mo - n_{Mo-1}) + (n¿¿Mo - n_{Mo+1})¿¿}$$

    Где x_{Mo}- нижняя граница модального интервала

    h-ширина интервала

    n_{Mo}- частота модального интервала

    Модальным считается интервал, которому соответствует наибольшая частота
- Медианой Me вариационного (статистического) ряда называется значение признака, приходящееся на середину ранжированного ряда наблюдений.

- Для дискретного вариационного ряда с нечетным числом членов медиана равна серединному варианту, а для ряда с четным числом членов, полусумме двух серединных вариантов, т.е.
  Если n=2k+1, то медиана Me=xk+1
  Если n=2k, то медиана Me=(xk+xk+1)/2
- Для интервального ряда медиана определяется по формуле:

$$Me = x_{Me} + \frac{h * \frac{n}{2} - S_{Me-1}}{n_{Me}}$$

x_{Me}-нижняя граница медианного интервала

h-ширина интервала

n_{Me}- частота медианного интервала

n-объем выборки

S_{Me-1}- сумма частот (накопленная частота) до медианного интервала

Медианным считается интервал, которому принадлежит значение признака с номером n/2 (если n – четное) или (n+1)/2 (если n-нечетное)

II Показатели вариации признака

- Выборочная дисперсия Dвыб – это среднее арифметическое квадратов отклонений значений признака от выборочной средней:
  - Простая: $D_{выб}(X) = \frac{\sum_{i=1}^{n} \left(x_i - \overline{x_{выб}}\right)^2}{n} = \overline{x^2} - \overline{x_{выб}}^2 = \frac{\sum_{i=1}^{n} x_i^2}{n} - \overline{x_{выб}}^2$
  - Взвешенная: $D_{выб}(X) = \frac{\sum_{i=1}^{k} \left(x_i - \overline{x_{выб}}\right)^2 * ¿}{\sum_{i=1}^{k} ¿} = \overline{x^2} - \overline{x_{выб}}^2 = \frac{\sum_{i=1}^{n} x_i^2 * ¿}{\sum_{i=1}^{k} ¿} - \overline{x_{выб}}^2$
- Выборочное среднее квадратическое отклонение (стандартное отклонение) – это арифметическое значение корня квадратного из выборочной дисперсии – рассчитывается по формуле: $\sigma_{выб} = \sqrt{D_{выб}}$
- Размах вариации R – это число, равное разности между наибольшим и наименьшим вариантами ряда, т.е. $R = x_{max} - x_{min}$
- Выборочный коэффициент вариации равен процентному отношению выборочного СКО к выборочной средней, т.е. $V(X) = \frac{\sigma_{выб}}{\overline{x_{выб}}} * 100\%$
- Выборочный начальный момент порядка k:
  $\widehat{a_k} = \frac{1}{n} \sum x_i^k$ или $\widehat{a_k} = \frac{1}{n} \sum x_i^k * n_i$
- Выборочный центральный момент порядка k:
  $\widetilde{\mu_k} = \frac{1}{n} \sum (x_i - \overline{x_{выб}})^k$ или $\widetilde{\mu_k} = \frac{1}{n} \sum (x_i - \overline{x_{выб}})^k * ¿$

Ш показатели формы распределения признака:

- Квантили (ранговые характеристики): q-квартили; d- децили; p-перцентили или процентили. – это значения признака, которые делят ранжированный ряд соответственно на 4, 10, 100 равных частей

- Асимметрия А. Показывает различия в вариации значений признака по одну и другую сторону от средней. $A_{выб}=\frac{\widetilde{\mu}_3}{\sigma^3_{выб}}$
- Эксцесс Е. Это показатель островершинности или плосковеришнности симметричного распределения по сравнению с нормальным распределением. $E_{выб}=\frac{\widetilde{\mu}_4}{\sigma^4_{выб}}-3$
- Асимметрия и эксцесс определяются для интервальных статистических рядов.

9. Понятие статистической оценки параметров распределения. Виды стат. оценок.

Статистическое оценивание – это определение приближенного значения неизвестного параметра генеральной совокупности по результатам наблюдения.

Параметр θ – это числовая характеристика генеральной совокупности.

Статистической оценкой ($\tilde{\theta}$) параметра тета называется его приближенное значение, зависящее от данных выбора.

Виды статистических оценок:

- Точечные

  Точечной оценкой θ* параметра θ называется числовое значение этого параметра, полученное по выборке объема n
- Интервальные

  Интервальной оценкой параметра тета называется числовой интервал (θ1; θ2), который с определенной вероятностью накрывает (содержит) неизвестное значение параметра.

10. Свойства статистических оценок: несмещенность, состоятельность, эффективность.

- Несмещенность (оценка совпадает с параметром θ). Стат оценка параметра θ называется несмещенной, если ее мат ожидание (среднее значение) равно оцениваемому параметру при любом объеме выборки $M(\tilde{\theta})=\theta \forall n$. В противном случае мат ожидание не равно параметру θ оценка называется смещенной. В этом случае разность между мат ожиданием и параметром называют смещением или систематической ошибкой оценивания.

  Св-во несмещенности является желательным, но не обязательным.
- Эффективность – оценка обладает наименьшей степенью случайных отклонений от параметра θ.

  Несмещенная оценка параметра θ называется эффективной, если она имеет наименьшую дисперсию среди всех возможных несмещенных оценок параметра, т.е. если ее дисперсия минимальна при заданном объеме выборки. Эффективность определяется отношением:

  $$e=\frac{\sigma^2_{\tilde{\theta}_э}}{\sigma^2_{\tilde{\theta}}}$$

  Где в числителе дисперсия эффективной оценки, а в знаменателе – дисперсия данной оценки.

  Чем ближе е к 1, тем эффективней оценка.

  Требование эффективности является желательным, но не обязательным.
- Состоятельность – оценка стремится к параметру θ с ростом объема выборки (n).

  Оценка параметра называется состоятельной, если она удовлетворяет ЗБЧ (т.е. при $n\to\infty$ сходится по вероятности к оцениваниваемому параметру)

  Другими словами, для любого положительного числа ε выполняется условие:

$$\forall \varepsilon>0;\ \lim_{n\to\infty} P\left(\left|\tilde{\theta}-\theta\right|<\varepsilon\right)=1$$

Свойства состоятельности является обязательным для любого правила оценивания.

11. Точечная оценка математического ожидания.

Пусть изучается СВ X с мат ожиданием a.

Несмещенной точечной оценкой мат ожидания M(X) $=\overline{x_{ген}}$ является выборочная средняя. $\overline{x_{ген}}\approx\overline{x_{выб}}$

12. Точечные оценки дисперсии и среднего квадратического отклонения.

$D_{ген}(X)=\sigma^2_{ген}(X)$ – генеральная дисперсия имеет 2 точечные оценки:

- $D_{выб}(X)=\sigma^2_{выб}(X)$- выборочная дисперсия, которая является смещенной оценкой
- S^2 – ¿исправленная выборочная дисперсия (несмещенная)

$$S^2=D_{выб}\frac{(X)*n}{n-1}\ (множитель\ Бесселя)$$

$\sigma_{ген}$ имеет 2 точечные оценки:

- $\sigma_{выб}-выб\ СКО$(стандартное отклонение) – смещенная $\sigma_{выб}=\sqrt{\sigma^2_{выб}}$
- S- исправленное выборочное СКО $S=\sqrt{S^2}$ (несмещенная)

13. Точечная оценка генеральной доли.

Пусть генеральная совокупность содержит N элементов, из которых M обладают некоторым признаком A. Тогда доля единиц, обладающих признаком A в генеральной доли обозначается $p=\frac{M}{N}$, где N- объем генеральной совокупности, M – число элементов в генеральной совокупности с признаком A.

Пусть n – объем выборки; m – число элементов с признаком A в выборке. Тогда величина $\omega=\frac{m}{n}$ называется выборочной долей.

Генеральную долю можно рассматривать как вероятность появления события A в биномиальном законе распределения СВ X, а выборочную долю – как относительную частоту этого события в n независимых испытаниях.

Несмещенной и состоятельной точечной оценкой генеральной доли является выборочная доля ω, т.е. $\hat{p}=\omega=\frac{m}{n}$

14. Методы нахождения точечных оценок: метод моментов, метод максимального правдоподобия.

Наиболее распространенные методы:

- Метод моментов (ММ)
- Метод максимального правдоподобия (ММП)
- Метод наименьших квадратов (МНК)

Метод моментов – это приравнивание теоретических моментов распределения (т.е. начальных αk или центральных μk) СВ X к соответствующим эмпирическим (выборочным) моментам, найденным по выборке, с последующим решением полученного уравнения (системы уравнений).

I Если теоретическое распределение задается одним параметрам, то для нахождения точечной оценки неизвестного параметра необходимо решить одно уравнение вида:

$$M(X)=\overline{x_{выб}}$$

II Если теоретическое распределение задается двумя параметрами, то для их оценки необходимо решить систему из двух уравнений вида:

$$\begin{cases} M(X)=\overline{x_{выб}} \\ D(X)=\sigma^2_{выб} \end{cases}$$

Достоинства и недостатки ММ:

- Неприменим, когда моменты генеральной совокупности нужного порядка не существуют
- Оценки, полученные данным методом, являются состоятельными, чего достаточно для решения многих задач
- Однако по эффективности эти оценки не являются наилучшими, их эффективность часто значительно меньше единицы
- ММ часто применяют на практике, т.к. он сводится к простым вычислениям.

Метод максимального правдоподобия (ММП).

В качестве оценки неизвестного параметра θ СВ X выбирается то его значение, при котором полученное значение выборки имеет наибольшую вероятность или плотность вероятности (для дискретных и непрерывных СВ соответственно).

Использование метода сводится к нахождению максимума функции одной или нескольких переменных.

Оценкой наибольшего правдоподобия параметра θ является такое его значение θ*, при котором функция правдоподобия L имеет максимум.

Алгоритм применения ММП:

1) По данной выборки {xn} построить функцию правдоподобия L(x1,x2…xn;θ)
2) Найти натуральный логарифм функции правдоподобия (упростить)
3) Найти первую производную LnL
4) Решить уравнение правдоподобия (LnL)'=0
5) Найти критические точки θ* -крит
6) Убедиться, что найденная точка θ* является точкой максимума, воспользовавшись достаточным условием максимума:
    a. Найти вторую производную (LnL)"
    b. Если (LnL)" | θ*<0 , то θ* - точка максимума.

Таким образом найденная точка θ* будет являться оценкой максимального правдоподобия параметра θ.

Оценки, полученные ММП являются состоятельными и эффективными, но не всегда несмещенными.

Недостатки ММП:

- Решение уравнения правдоподобия бывает трудным
- Нужно знать закон распределения СВ X, что на практике часто бывает невозможным

15. Понятие доверительной вероятности и доверительного интервала. Точность оценки.

Задача интервального оценивания – по данным выборки построить числовой интервал (θ1; θ2), относительно которого с заранее выбранной вероятностью γ можно сказать, что внутри этого интервала находится оцениваемый параметр.

Интервальной оценкой параметра θ называется числовой интервал (θ1; θ2), который с заданной вероятностью гамма накрывает (содержит) неизвестное значение параметра θ т.е. для которого выполняется условие: $P(\theta 1<\theta<\theta 2)=\gamma$

Величина γ – есть вероятность события, что параметр θ отличается от своей точечной оценки на величину, не превосходящую некоторого положительного числа δ

$\gamma=P(|\theta-\theta^{¿}|<\delta)$, где δ- точность оценки

Интервал (θ1; θ2) называется доверительным интервалом или интервальной оценкой, а вероятность γ – доверительной вероятностью, уровнем доверия или надежностью оценки.

Число α=1-γ – уровень значимости:

- Значение α(γ) выбирают заранее
- Выбор зависит от конкретно решаемой задачи
- Обычно выбирают: γ= 0,9; 0,95; 0,99; 0,999
- Или α=0,1; 0,05; 0,01; 0,001
- Ширина интервала h=θ2- θ1 существенно зависит от объема выборки n (уменьшается с ростом n) и от значения доверительной вероятности (увеличивается с приближением γ к единице)

16. Основные законы распределения математической статистики.

Распределение Хи-квадрат (Пирсона) с k=n степенями свободы называют распределение суммы квадратов независимых СВ, распределенных по стандартному нормальному закону, т.е.

$Z=Y_1^2+Y_2^2+...+Y_n^2$, где Yi – независимы и имеют одно и то же нормальное стандартное распределение, т.е. Yi~N(0,1); M(Yi)=0, D(Yi)=1

Число слагаемых называется числом степеней свободы распределения (это параметр формы кривой χ^2-распределения)

Обозначается Z~$\chi^2(n)$

СВ, распределенная по закону χ^2 принимает только неотрицательные значения, т.е. Z≥0

Кривая χ^2-распределения одновершинная, асимметричная. С ростом n вершина кривой сдвигается вправо от начала координат и χ^2- распределение медленно приближается к нормальному.

Числовые характеристики: M(χ^2¿=n; D(χ^2¿=2n

Распределение χ^2 используется при интервальном оценивании дисперсии, проверке статистических гипотез согласия, однородности, независимости, при проверке значимости коэффициента корреляции и в других задачах статистического анализа данных.

Распределение Стьюдента или t-распределение с k степенями свободы называется распределение СВ $Z=\frac{X}{\sqrt{\frac{Y}{k}}}$, где X- СВ, распределенная по стандартному нормальному закону (X~N(0,1)), а Y имеет распределение Хи-квадрат с k степенями свободы, т.е. Y~$\chi^2(k)$. Обозначение Z~t(k). СВ, распределенная по закону Стьюдента принимает любые значения.

Распределение Стьюдента симметрично относительно 0: если X~t(k), то и (-X) ~t(k).

Числовые характеристики: M(t)=Mo(t)=Me(t)=0; D(t)=$\frac{k}{k-2}$, k>2

Кривая распределения Стьюдента по сравнению с кривой нормального распределения является более пологой. При k→∞, t-распределение быстро приближается к нормальному, при k>30 они уже практически неразличимы.

Распределение Стьюдента применяют при интервальном оценивании мат ожидания и других параметров, при проверке гипотез о значениях мат ожиданий, оценке значимости коэф-ов регрессии, гипотез об однородности выборок и тд.

Распределением Фишера с k1 и k2 степенями свободы называется распределение СВ $Z=\frac{X}{k_1}:\frac{Y}{k_2}=\frac{X*k_2}{Y*k_1}$, где X и Y – независимые СВ распределенные по закону Пирсона, т.е. X~ $\chi^2(k1)$; Y~$\chi^2(k2)$

Обозначение Z~F(k1,k2). СВ, распределенная по закону Фишера может принимать только неотрицательные значения, т.е. $Z\geq 0$

Числовые характеристики:

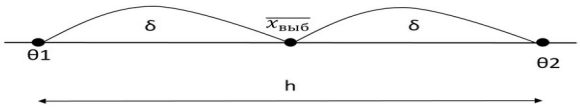
$$M(F)=\frac{k_2}{k_2-2},\ k_2>2$$

$$D(F)=\frac{2k_2^2(k_1+k_2-2)}{k_1(k_2-2)^2(k_2-4)},\ k_2>4$$

С ростом числа степеней свободы распределение Фишера приближается к нормальному. Применяется при проверке гипотез о равенстве дисперсий, о статистической значимости уравнения регрессии.

17. Интервальная оценка генеральной средней (математического ожидания).

В качестве интервальной оценки мат ожидания выступает точечная оценка – выборочная средняя – относительно которой строится симметричный интервал.



Правила построения доверительного интервала зависят от того, известна или неизвестна генеральная дисперсия $\sigma^2_{ген}$

- Доверительный интервал с заданной надежностью γ неизвестного математического ожидания с известной генеральной дисперсией имеет вид:

  $$\overline{x_{выб}}-\frac{\sigma_{ген}}{\sqrt{n}}z<a<\overline{x_{выб}}+\frac{\sigma_{ген}}{\sqrt{n}}z$$

  Где Z – квантиль стандартного нормального распределения уровня 1-α/2

  В Excel: =НОРМ.СТ.ОБР((1+γ)/2)

  Число $\delta=\frac{\sigma_{ген}}{\sqrt{n}}z$ – точность оценки мат ожидания. Тогда доверительный интервал для генеральной средней можно записать в виде: $(\theta_1;\theta_2)=(\overline{x_{выб}}-\delta;\overline{x_{выб}}+\delta)$

  Ширина интервала $h=\theta_2-\theta_1=2\delta$

- Доверительный интервал с заданной надежностью γ неизвестного математического ожидания с неизвестной генеральной дисперсией имеет вид:

$\overline{x_{выб}} - \frac{S}{\sqrt{n}} t_\alpha < a < \overline{x_{выб}} + \frac{S}{\sqrt{n}} t_\alpha$, где t_α – ¿квантиль распределения Стьюдента (для двусторонней области) соответствующий k=n-1 степеням свободы и уровню значимости α

В Excel: = СТЬЮДЕНТ.ОБР.2Х(α;k=n-1)

Число $\delta = \frac{S}{\sqrt{n}} t_\alpha$ – точность оценки мат ожидания.

18. Интервальная оценка генеральной дисперсии и генерального СКО.

Основой интервальной оценки генеральной дисперсии является статистика S^2 (Исправленная выборочная дисперсия или исправленное выборочное СКО – S)

Интервал в отличии от генеральной средней для генеральной дисперсии строится несимметричный.

Доверительный интервал с заданной надежностью γ для генеральной дисперсии имеет вид:

$$\frac{(n-1)*S^2}{\chi_2^2} < \sigma_{ген}^2 < \frac{(n-1)*S^2}{\chi_1^2}$$

Где критические точки определяются по таблице распределения Пирсона:

$$\chi_1^2 = \chi^2\left(\frac{(1-\gamma)}{2}; n-1\right)$$

$$\chi_2^2 = \chi^2\left(\frac{(1+\gamma)}{2}; n-1\right)$$

Замечание: значения критических точек распределения можно найти с помощью функций Excel:

χ_1^2=ХИ2.ОБР($\frac{(1-\gamma)}{2}; n-1$)

χ_2^2=ХИ2.ОБР($\frac{(1+\gamma)}{2}; n-1$)

Доверительный интервал для генерального СКО имеет вид:

$$S*\sqrt{\frac{n-1}{\chi_2^2}} < \sigma_{ген}^2 < S*\sqrt{\frac{n-1}{\chi_1^2}}$$

19. Интервальная оценка генеральной доли.

Доверительный интервал для генеральной доли с заданной надежностью γ строится симметрично относительно выборочной доли и имеет вид: $\omega - \delta < p < \omega + \delta$

Где $\omega = \frac{m}{n}$ – выборочная доля

$\delta = z*\sqrt{\frac{\omega(\omega-1)}{n}}$ точность оценки генеральной доли

Z - квантиль стандартного нормального распределения уровня $1-\frac{\alpha}{2}$

Z=НОРМ.СТ.ОБР((1+ γ)/2)

20. Понятие статистической гипотезы. Параметрические и непараметрические гипотезы.

Статистическая гипотеза – это некоторое высказывание относительно генеральной совокупности, проверяемое по выборочным данным.
Статистическая гипотеза – это любое предположение о виде неизвестного распределения или о параметрах известного закона распределения.
Статистические гипотезы:

- Параметрические
  - Простые
  - Сложные
- Непараметрические

Параметрические гипотезы – это утверждения о значении параметров распределения известного вида:

- Простые – утверждения о том, что значение неизвестного параметра генеральной совокупности равно одному заданному числу
- Сложные – гипотезы, не являющиеся простыми

Непараметрические гипотезы – это утверждения о виде неизвестного распределения

21. Нулевая и альтернативная гипотеза. Типы альтернативных гипотез.

Выдвинутая гипотеза называется нулевой (основной). H0 – нулевая гипотеза. По отношению к основной гипотезе можно сформулировать альтернативную (конкурирующую) противоречащую ей, т.е. гипотезу противоположную H0. H1 – альтернативная гипотеза.
H0 и H1 – это два предположения, из которых в результате статистической проверки должно быть выбрано только одно.
Проверить статистическую гипотезу значит установить, согласуются ли данные, полученные из выборки с этой гипотезой.
Статистическими методами гипотезу можно только опровергнуть или не опровергнуть, но не доказать!!!

Примеры параметрических гипотез:

- H0: мат ожидание СВ X равно 2 (нулевая, основная)
- H1: $a \neq 2$ – двусторонняя сложная гипотеза
- H1: $a > 2$ – правосторонняя сложная гипотеза
- H1: $a < 2$ – левосторонняя сложная гипотеза
- H1: $a = 1{,}9$ – простая левосторонняя гипотеза
- H1: $a = 2{,}1$ – простая правосторонняя гипотеза

22. Задача проверки статистических гипотез. Понятие статистического критерия.

Сопоставление высказанной гипотезы о генеральной совокупности с имеющимися выборочными данными, сопровождаемое количественной оценкой степени достоверности получаемого вывода и осуществляемое с помощью того или иного статистического критерия называется проверкой статистических гипотез.
Статистический критерий K (или стат тест) – это правило (формула), по которому определяется мера расхождения результатов выборочного наблюдения с высказанной гипотезой H0, т.е. согласно которому принимается решение, принимать или отклонять нулевую гипотезу.

Основу критерия представляет специально составленная выборочная характеристика (СВ или статистика) $\theta_n = \theta(X_1, X_2, \dots X_n)$- точечное или приближенное распределение которой неизвестно.

23. Наблюдаемое и критическое значения статистического критерия.

Значение критерия, рассчитываемое по специальным правилам (по формулам) на основании выборочных данных называют наблюдаемым значением критерия (Кнабл)

Множество значений критерия К:

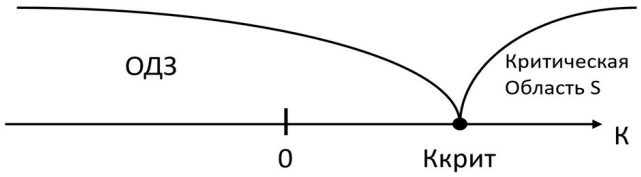
- Область допустимых значений (область принятия нулевой гипотезы) $\overline{S}$ – совокупность всех значений критерия К, при которых нулевая гипотеза Н0 не отклоняется
- Критическая область S – совокупность всех значений критерия К, при которых нулевая гипотеза отклоняется в пользу конкурирующей Н1.

Значения критерия, разделяющие области S и $\overline{S}$, определяемые на заданном уровне значимости α по таблицам распределения (инструментальными средствами) СВ К, выбранной в качестве критерия, называют критическими точками (Ккрит)

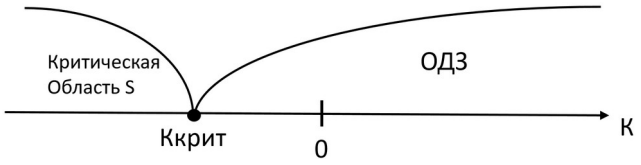
24. Критическая область и ее типы. Область принятия гипотезы.

Вид критической области зависит от того, какая гипотеза выдвинута в качестве альтернативной:

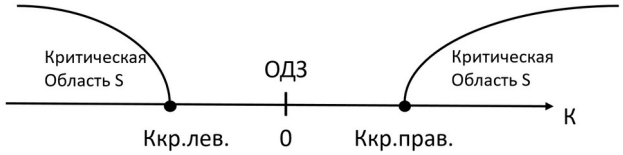
- Если конкурирующая гипотеза правосторонняя (Н1: a>2), то и критическая область S правосторонняя. В этом случае будет одна критическая точка, которая принимает положительные значения (Ккрит.прав.>0)



- Если конкурирующая гипотеза левосторонняя (Н1: a<2), то и критическая область S левосторонняя. В этом случае будет одна критическая точка, принимающая отрицательные значения. (Ккр.лев.<0)



- Если конкурирующая область двусторонняя (Н1: a≠2), то и критическая область S будет двусторонняя. В этом случае будет 2 критических точки (Ккр.лев<0 и Ккр.прав>0), которые будут симметричны относительно нуля.



25. Основной принцип проверки статистических гипотез.

Основной принцип проверки статистических гипотез:

- Если наблюдаемое значение критерия (Кнабл) принадлежит критической области S, то нулевая гипотеза отклоняется в пользу конкурирующей гипотезы Н1
- Если наблюдаемое значение критерия (Кнабл) принадлежит ОДЗ, то нулевую гипотезу Н0 нельзя отклонять.

26. Общая схема проверки статистических гипотез.

1) Располагая выборкой x1,x2…xn сформулировать основную (H0) и альтернативную (H1) гипотезы
2) Выбрать уровень значимости α для проведения проверки
3) По виду H0 выбрать статистический критерий для ее проверки
4) На основании данных выборки по специальному алгоритму (формуле) найти наблюдаемое значение критерия Кнабл
5) По таблицам (формулам Excel) распределения СВ К, выбранной в качестве статистического критерия, найти критическое значение (критическую точку или точки)
6) Исходя из типа конкурирующей гипотезы H1 определить тип критической области S
7) Посмотреть, какой области (S или $\overline{S}$) принадлежит наблюдаемое значение критерия
8) Сделать вывод о принятии или отклонении нулевой гипотезы.

Замечание: даже в том случае, если H0 нельзя отклонить, это еще не значит, что данное предположение о генеральной совокупности является единственным верным, просто ему не противоречат имеющиеся выборочные данные. Таким свойством могут обладать и другие гипотезы.

27. Ошибки первого и второго рода. Мощность критерия.

При проверке гипотезы могут быть приняты неправильные решения, т.е. могут быть допущены ошибки первого и второго рода.

Ошибкой первого рода называется ошибка, возникающая, когда нулевая гипотеза отклоняется в то время как в действительности в генеральной совокупности она является справедливой. Вероятность совершить ошибку первого рода называется уровнем значимости (или размером критерия) и обозначается α.

Ошибкой второго рода называется ошибка, возникающая, когда H0 принимается, в то время как на самом деле в генеральной совокупности она является ошибочной, а справедлива альтернативная H1.

Вероятность допустить ошибку второго рода называется β

Вероятность 1- β, т.е. вероятность не допустить ошибку второго рода называют мощностью критерия (или функцией мощности)

Выбор статистического критерия и критической области осуществляют таким образом, чтобы мощность критерия была максимальной.

Критерий называют наиболее мощным, если из всех критериев с заданным уровнем значимости α, он имеет наибольшую мощность (наименьшее значение β)

Риски при проверке гипотез

Нулевая гипотеза H0	Результаты решения относительно нулевой гипотезы	
	Отклонена	Принята
Верна	Ошибка первого рода P(H1/H0)=α (уровень значимости)	Правильное решение P(H0/H0)=1-α (надежность)
Неверна	Правильное решение P(H1/H1)=1-β (мощность критерия)	Ошибка второго рода P(H0/H1)=β

28. Наблюдаемый уровень значимости (p-значение/p-value)

Наблюдаемым уровнем значимости (или p-значением или p-value) гипотезы H0, проверяемой по выборке x1,x2…xn с помощью критерия K(x1,x2…xn) называется наименьшее значение α, при котором нулевая гипотеза отклоняется

$$p(x1, x2 \dots xn) = min\{\alpha : K(x1, x2 \dots xn) \in S$$

Ещё одна интерпретация p-значения – это вероятность, с которой (при условии истинности H0) могла бы реализоваться полученная выборка или любая другая выборка с еще менее вероятным наблюдаемым значением статистики K.

СВ p(x1,x2…xn) имеет равномерное распределение.

- Если p(x1,x2…xn)<α, то есть основания отклонять H0
- Если p(x1,x2…xn)≥α, то нет оснований отклонять H0

Поэтому на практике, чем меньше p-значение, тем меньше вероятность ошибиться, отклонив нулевую гипотезу и тем выше уверенность в том, что необходимо отвергнуть H0.

29. Гипотезы о равенстве средних 2-х нормально распределенных генеральных совокупностей.

Случай 1: генеральные дисперсии известны (z-тест)

1) Найдем наблюдаемое значение критерия. Найдем точечные оценки мат ожидания (генеральные средние). При достаточно больших объемах выборок (≥30) выборочные средние имеют приближенно нормальный закон распределения, поэтом при выполнении H0 статистика z имеет стандартный нормальный закон распределения, т.е. формула (1):

$$Z = \frac{\overline{x_{выб}} - \overline{y_{выб}}}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}} \quad N(0;1)$$

   По формуле (1) находим наблюдаемое значение критерия.

2) Найдем критические точки и критическую область S. Критическая область зависит от типа альтернативной гипотезы H1.
   a. H1 – двусторон. S-двусторон. Найдем 2 симметричные критические точки:
      Zкр.пр. – квантиль станд норм расп. Уровня 1-α/2 (=НОРМ.СТ.ОБР(1-α/2))
      Zкр.лев. – квантиль станд норм расп уровня α/2(=-Zкр.пр.)
      S=(-∞;Zкр.лев.)U(Zкр.пр.;+∞)
   b. H1- правосторон S-правосторон. Найдем одну критическую точку (правую):
      Zкр.пр. – квантиль станд норм расп уровня 1- α (=НОРМ.СТ.ОБР(1- α))
      S=(Zкр.пр.; +∞)

3) Принятие решения относительно нулевой гипотезы: Если znабл принадлежит критической области, то H0 отклоняется в пользу H1. Иначе нет оснований отклонять H0.

Замечание 1: Если в условии даны несгруппированные ряды, то Z-тест в Excel можно реализовать с помощью инструмента в пакете анализа «Двухвыборочный Z-тест для средних».

Замечание 2: Этот инструмент дает значение p-value только для односторонней H1, поэтому если H1-двусторон, то полученное p-value надо умножить на 2.

Случай 2: генеральные дисперсии неизвестны (t-тест – критерий Стьюдента)

1) Найдем несмещенные точечные оценки генеральных средних и генеральных СКО.

$$\hat{a}_x = \overline{x_{выб}}$$
$$\hat{a}_y = \overline{y_{выб}}$$
$$\hat{\sigma}_x = S_x$$
$$\hat{\sigma}_y = S_y$$

При условии выполнения H0, статистика t имеет распределение Стьюдента с $k=n_x+n_y-2$ степенями свободы, т.е. формула (2):

$$t = \frac{\overline{x_{выб}} - \overline{y_{выб}}}{S * \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}} \quad t(k)$$

$S = \sqrt{\frac{(n_x-1)*S_x^2 + (n_y-1)*S_y^2}{n_x+n_y-2}}$ – объединенная оценка СКО

По формуле (2) находим t набл

2) Найдем критические точки и критическую область:
   a. H1 – двусторон. S- двусторон.
      Найдем 2 симметричные критические точки.
      $t_{кр.пр.}$- квантиль распределения Стьюдента уровня 1-α/2 (=СТЬЮДЕНТ.ОБР(1-α/2;k) =СТЬЮДЕНТ.ОБР.2Х(α;k))
      $t_{кр.лев.}$- квантиль распределения Стьюдента уровня α/2 = $-t_{кр.пр.}$

   $S=(-\infty; t_{кр.лев.})U(t_{кр.пр.}; +\infty)$

   b. H1-правосторон S-правосторон
      Найдем одну критическую точку (правую)
      $t_{кр.пр.}$- квантиль распределения Стьюдента уровня 1-α (=СТЬЮДЕНТ.ОБР(1-α;k) =СТЬЮДЕНТ.ОБР.ПХ(α;k)
      $S=(t_{кр.пр.}; +\infty)$

3) Принятие решения относительно нулевой гипотезы. Если наблюдаемое значение t принадлежит критической области, то H0 отклоняется в пользу H1 на уровне значимости альфа, иначе нет оснований отклонять H0.

Замечание 1: Описанный способ проверки с помощью объединенного двухвыборочного t-теста можно проводить только когда генеральные дисперсии совокупностей X и У можно считать равными, поэтому перед проверкой H0 нужно сначала проверить гипотезу о равенстве генеральных дисперсий с помощью критерия Фишера (F-теста).

Замечание 2: Если в условии даны несгруппированные выборки, то проверку можно осуществить с помощью инструментов Excel:

- Если Генеральные дисперсии можно считать равными, то используем «двухвыборочный t-тест с одинаковыми дисперсиями
- Если генеральные дисперсии нельзя считать равными, то используем «двухвыборочный t-тест с различными дисперсиями» (!!!p-значение в этом случае считается неверно)

Замечание 3: Если не проводить проверку гипотезы о равенстве генеральных дисперсий (неизвестно равны они или нет), то t-тест можно реализовать с помощью функций Excel.

P-value:

- =СТЬЮДЕНТ.ТЕСТ(МассивХ ; Массив У; Хвосты (1, если Н1 – односторонняя, 2, если двусторонняя); Тип (3))
- =ТТЕСТ(аргументы те же)

Если р-значение меньше альфа, то Н0 отклоняется в пользу Н1, иначе нет оснований отклонять Н0.

30. Гипотеза о равенстве дисперсий 2-х нормально распределенных генеральных совокупностей.

1) Найдем наблюдаемое значение критерия:
   Найдем точечные несмещенные оценки генеральных дисперсий (S^2)
   Если Н0 выполняется, то F-статистика имеет распределение Фишера С k1,k2 степенями свободы, т.е. по формуле (3):

$$F = \frac{S^2_{большая}}{S^2_{меньшая}} \quad F(k1, k2)$$

   Где k1=n1-1 (n1- объем выборки с большей исправленной выб дисперсией), k2=n2-1 (n2 – объем выборки с меньшей исп выб дисп)
   По формуле (3) найдем наблюдаемое значение критерия.
2) Найдем критические точки и критическую область
   - Если Н1 двусторон, то S- двусторон.
     F≥0, найдем 2 критические точки.
     Fкр.лев. – квантиль распределения Фишера уровня α/2 (=F.ОБР(α/2;k1;k2) **<1**
     Fкр.пр – квантиль распределения Фишера уровня 1- α/2 (=F.ОБР(1- α/2;k1;k2) **>1**
     S=(0;Fкр.лев)U(Fкр.пр.;+∞)
   - Если Н1 правосторон, то S-правосторон
     F≥0, найдем Fкр пр
     Fкр.пр – квантиль распределения Фишера уровня 1-α (=F.ОБР(1- α;k1;k2); =F.ОБР.ПХ(α;k1;k2))
     S=(Fкр.пр;+∞)
3) Принятие решения относительно Н0: Если наблюдаемое значение принадлежит критической области, то Н0 отклоняется в пользу Н1, иначе нет оснований отклонять Н0.

Замечание: По несгруппированным выборкам F- тест можно провести следующими способами:

- С помощью инструмента надстройки анализа данных «двухвыборочный F-тест для дисперсий»:
  - Первый массив нужно ввести массив с наибольшей исп выб дисп
  - Если Н1 односторонняя, то в строку «Альфа» вводим значение α, если двусторонняя, то α/2
  - Этот инструмент считает р-значение для односторонней Н1, если Н1-двусторонняя, то р-значение надо умножить на 2
- С помощью функции =F.ТЕСТ(Массив1;Массив2) – считает р-значение только для двусторонней Н1, если Н1 односторонняя, то значение надо разделить на 2.

31. Гипотеза о равенстве генеральный долей признака в двух совокупностях.

Сравнение вероятностей успеха в двух сериях испытания Бернулли (z-тест)

Имеется 2 выборки, распределенные по биномиальному закону с параметрами (ni; pi), где pi-вероятность успеха в одном испытании Бернулли.

Проверка:

1) Найдем наблюдаемое значение критерия.

   По выборкам A:n1;k1, где k1 – число элементов с признаком и Б:n2;k2. n1 и n2 ≥30. Найдем несмещенные точечные оценки генеральных долей, т.е. выборочные доли:

   $p_1^{¿} = \omega_1 = \frac{k_1}{n_1}$

   $p_2^{¿} = \omega_2 = \frac{k_2}{n_2}$

   При достаточно больших объемах выборок выборочные доли имеют приближенно нормальный закон распределения. При условии, что H0 выполняется Z-статистика имеет стандартное нормальное распределение, т.е. по формуле (4):

   $$Z = \omega_1 - \omega_2 - \frac{1}{2¿¿}¿$$

   $\omega = \frac{k_1 + k_2}{n_1 + n_2}$ – общая доля из двух выборок

   По формуле 4 *находим Z набл.*

2) Найдем критические точки и критическую область

   - Если H1 –двусторон, S- двусторон,
     Найдем 2 симметричные критические точки
     Zкр.пр. – квантиль станд норм расп уровня 1-α/2 (=НОРМ.СТ.ОБР(1-α/2))
     Zкр.лев. – квантиль станд норм расп уровня α/2 (-Zкр.пр.)
     S(-∞;Zкр.лев)U(Zкр.пр.;+∞)
   - Если H1 – правосторон, S – правосторон
     Найдем Zкр пр
     Zкр. Пр. – квантиль станд норм рас уровня 1-α (=НОРМ.СТ.ОБР(1-α))
     S=(Zкр.пр.;+∞)

3) Принятие решения

Замечание: Z-тест можно использовать только при больших объемах выборки, т.е. когда выборки по объему репрезентативны (представительны):

- $n_1 * \omega_1 \geq 5$
- $n_1 * (1 - \omega¿¿1) \geq 5¿$
- $n_2 * \omega_2 \geq 5$
- $n_2 * (1 - \omega¿¿2) \geq 5¿$

Должны выполняться все 4 условия

32. Гипотезы о числовых значениях параметров распределения.

33. Суть и схема применения критерия согласия Пирсона.