Математическая статистика для чайников

Консультация

2020 - 2021 учебный год

Подготовила Андреева Софья,
БПИ183

Содержание

Статистические оценки

Возьмем некую выборку $X_1, \ldots, X_n$, в которой все случайные величины X_i имеют распределение, заданное функцией $F(x, \theta_1, \ldots, \theta_k)$. Здесь $\theta_1, \ldots, \theta_k$ – это некоторые ***параметры распределения***, которые нам неизвестны, при этом их может как несколько, так и один.

К примеру, может быть известно, что $X_1, \ldots, X_n$ имеют нормальное распределение, но его параметры неизвестны, то есть $X_1, \ldots, X_n \sim N(\theta_1, \theta_2^2)$ (на всякий случай напомню, что обычно нормальное распределение имеет параметры μ и $\sigma^2 : N(\mu, \sigma^2)$). А может оказаться так, что они имеют тоже нормальное распределение, но неизвестен только параметр $\mu : X_1, \ldots, X_n \sim N(\theta_1, 5^2)$.

Так как некоторые параметры неизвестны, но известна выборка, хочется хотя бы примерно оценить, какие значения могут принимать эти параметры. И один из способов сделать это – придумать точечную оценку.

Оценкой одного параметра θ будем называть функцию от $X_1, \ldots, X_n$, не зависящую от θ. Обозначается она так: $\hat{\theta} = f(X_1, \ldots, X_n)$. Вообще говоря, это может быть абсолютно любая функция.

Например, захотел кто-то оценить математическое ожидание на выборке из 10 случайных величин, и сказал, что придумал функцию $5X_2 - 7X_6 + X_9$. Это тоже является оценкой, но тут уже стоит задаться вопросом, насколько хорошая эта оценка. И вот для того, чтобы понять, придумана хорошая оценка или плохая, существуют несколько свойств.

Свойства оценок

1. Несмещенность
Оценка $\hat{\theta}$ называется ***несмещенной***, если ее математическое ожидание равно параметру θ

$$E(\hat{\theta}) = \theta$$

Например, рассмотрим оценку математического ожидания, которая известна как ***выборочное среднее***: $\hat{m} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Посчитаем для нее математическое ожидание.

$$E\bar{X} = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} EX_i = \frac{1}{n} n EX_1 = m$$

(Немного пояснений по рассчетам: так как все X_i одинаково распределены, то у них всех будет одинаковое математическое ожидание, поэтому если просуммировать их все, получим $n \cdot m$)

Итак, посчитав все это, получили, что выборочное среднее является несмещенной оценкой для математического ожидания. А если немного изменить оценку, прибавив к ней некоторое число, то получится, что она уже станет смещенной. Например, если взять оценку $\hat{m} = 3 + \frac{1}{n}\sum_{i=1}^{n} X_i$, то ее математическое ожидание будет

$$E\hat{m} = E\left(3 + \frac{1}{n}\sum_{i=1}^{n} X_i\right) = E(3) + E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = 3 + m$$

Это не равно m, а значит оценка не является несмещенной.

Существует еще одна известная смещенная оценка для дисперсии, которую называют ***выборочной дисперсией***: $S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$. Ее математичекое ожидание равно $ES^2 = \frac{n-1}{n}\sigma^2$. Выписывать здесь подсчеты не буду, они долгие и сложные, поэтому если хотите, можете сами попробовать это посчитать. А также есть еще оценка для дисперсии, которая уже является несмещенной – это ***несмещенная выборочная дисперсия***: $\tilde{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$. Ее математическое ожидание равно $E\tilde{S}^2 = \sigma^2$.

2. Асимптотическая несмещенность
Оценка $\hat{\theta}$ называется ***асимптотически несмещенной***, если ее математическое ожидание стремится к параметру θ, когда объем выборки стремится к бесконечности.

$$\lim_{n\to\infty} E(\hat{\theta}) = \theta$$

Рассмотрим выборочную дисперсию, выписанную немного выше. Она смещенная, но если n стремится к бесконечности, то $ES^2 = \frac{n-1}{n}\sigma^2 \to \sigma^2$, а значит, она является асимптотически несмещенной.

3. Состоятельность
Оценка $\hat{\theta}$ называется ***состоятельной***, если она сходится по вероятности к θ.

$$\hat{\theta} \xrightarrow[n\to\infty]{P} \theta$$

Если вспомнить, что такое сходимость по вероятности, то получается, что

$$\forall \varepsilon > 0 \quad P(|\hat{\theta} - \theta| > \varepsilon) \xrightarrow[n\to\infty]{} 0$$

Или, если хочется, чтобы было меньше ε, а не больше, то можно использовать формулу

$$\forall \varepsilon > 0 \quad P(|\hat{\theta} - \theta| \leq \varepsilon) \xrightarrow[n\to\infty]{} 1$$

Если вспомнить здесь еще неравенство Чебышева, то можно получить достаточное условие состоятельности. По неравенству Чебышева $P(|\xi - E\xi| \leq \varepsilon) \geq 1 - \frac{D\xi}{\varepsilon^2}$. Подставим вместо ξ оценку $\hat{\theta}$:

$$P(|\hat{\theta} - E\hat{\theta}| \leq \varepsilon) \geq 1 - \frac{D\hat{\theta}}{\varepsilon^2}$$

И тогда очевидны два условия для состоятельности:
1. $\lim_{n\to\infty} E(\hat{\theta}) = \theta$
2. $\lim_{n\to\infty} D(\hat{\theta}) = 0$

Если все это выполняется, то при $n \to \infty$:

$$P(|\hat{\theta} - \theta| \leq \varepsilon) \geq 1$$

А так как вероятность не может больше чем 1, то получим наше первоначальное определение состоятельности

$$P(|\hat{\theta} - \theta| \leq \varepsilon) \xrightarrow[n\to\infty]{} 1$$

Эти два условия возможно не выводились на лекциях, поэтому в работах лучше прописывать, что использовалось неравенство Чебышева. Но во многих случаях удобнее использовать их, чем полностью считать вероятность. Также следует обратить внимание, что это условия достаточности, то есть возможны ситуации, когда оценка является состоятельной, но эти условия не

выполняются.

Рассмотрим выборочное среднее. Его математическое ожидание уже считалось выше, когда доказывалась его несмещенность. И, что понятно, выборочное среднее является и ассимптотически несмещенным, то есть выполняется первое условие. Посчитаем еще дисперсию:

$$D\bar{X} = D\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} DX_i = \frac{1}{n^2} nDX_1 = \frac{\sigma^2}{n} \xrightarrow[n\to\infty]{} 0$$

(Здесь использовалось то, что все величины в выборке независимы и имеют одинаковую дисперсию)

Значит, так как выполнились оба условия, то выборочное среднее является состоятельной оценкой. Также можно было просто воспользоваться законом больших чисел, но для этого его надо было вспомнить.

4. Сильная состоятельность
Оценка $\hat{\theta}$ называется ***сильно состоятельной***, если она сходится почти наверное к θ.

$$\hat{\theta} \xrightarrow[n\to\infty]{\text{п.н.}} \theta$$

Если вспомнить определение сходимости почти наверное, то

$$P(\omega : \lim_{n\to\infty} \hat{\theta}(\omega) = \theta(\omega)) = 1$$

Это свойство используется очень редко, даже можно сказать, что никогда, поэтому про него стоит немножко помнить и не использовать.

5. Эффективность
Оценка $\hat{\theta}$ называется ***эффективной***, если она несмещенная и ее дисперсия является наименьшей среди всех несмещенных оценок для этого параметра θ.

В основном на практике используются два свойства – несмещенность и состоятельность. Про остальные нужно просто не забывать, что это такое. И, собственно, если оценка является несмещенной и состоятельной, то это говорит о том, что эта оценка является достаточно хорошей.

Но как вообще придумывают оценки? Как придумать такую функцию, да еще чтобы выполнялись какие-то полезные свойства? Для этого существуют два известных метода.

Методы нахождения и построения оценок

1. Метод моментов
Рассморим начальные моменты $\mu_j = EX^j, j = \overline{1,k}$, где k – количество параметров, которые нужно оценить. X здесь – это случайная величина, порождающая выборку, то есть величина, у которой распределение такое же, как у всех случайных величин выборки. Понятно, что эти моменты будут зависеть от неизвестных θ_i. А теперь приравняем каждый момент соответствующему ему начальному выборочному моменту, которые равны $\hat{\mu}_j = \frac{1}{n}\sum_{i=1}^{n} X_i^j$. И получим следующую систему:

$$\begin{cases} \mu_1 = \hat{\mu}_1 \\ \vdots \\ \mu_k = \hat{\mu}_k \end{cases}$$

Если система решается и решается однозначно, то ее решение будет являться оценками параметров $\theta_1, \dots, \theta_k$.

И рассмотрим пример, как это используется. Пусть есть выборка из n случайных величин, которые имеют нормальное распределение с неивестными параметрами μ и σ: $X_1, \dots, X_n \sim N(\theta_1, \theta_2^2)$. Тогда составим систему для того, чтобы оценить θ_1 и θ_2.

$$\begin{cases} EX = \dfrac{1}{n}\sum\limits_{i=1}^{n} X_i \\ EX^2 = \dfrac{1}{n}\sum\limits_{i=1}^{n} X_i^2 \end{cases}$$

$EX = \theta_1$ из условия, также из условия известно $DX = \theta_2^2$. Тогда можно сказать, чему равно EX^2 из определения дисперсии. Так как $DX = EX^2 - (EX)^2$, то $EX^2 = DX + (EX)^2 = \theta_2^2 + \theta_1^2$. А также известно, чему равен первый выборочный момент – это выборочное среднее. Запишем все это в систему:

$$\begin{cases} \theta_1 = \bar{X} \\ \theta_2^2 + \theta_1^2 = \dfrac{1}{n}\sum\limits_{i=1}^{n} X_i^2 \end{cases} \Rightarrow \begin{cases} \theta_1 = \bar{X} \\ \theta_2^2 = \dfrac{1}{n}\sum\limits_{i=1}^{n} X_i^2 - \bar{X}^2 \end{cases}$$

Это и есть полученные оценки:

$$\begin{cases} \hat{\theta}_1 = \bar{X} \\ \hat{\theta}_2^2 = \dfrac{1}{n}\sum\limits_{i=1}^{n} X_i^2 - \bar{X}^2 \end{cases}$$

2. Метод максимального правдоподобия

Введем функцию правдоподобия

$$L(X_1, \dots, X_n, \theta_1, \dots, \theta_k) = \begin{cases} \prod\limits_{i=1}^{n} f(x_i, \theta_1, \dots, \theta_k), \text{если распределение непрерывное} \\ \prod\limits_{i=1}^{n} P(X = x_i, \theta_1, \dots, \theta_k), \text{если распределение дискретное} \end{cases}$$

Оценкой максимального правдоподобия называется максимум этой функции, то есть
$\hat{\theta} = \underset{\theta \in \mathbb{R}^k}{\operatorname{argmax}} L(X_1, \dots, X_n, \theta)$, где $\theta = (\theta_1, \dots, \theta_k)^T$

Так как функция правдоподобия и логарифмическая функция правдоподобия $(\ln L(X_1, \dots, X_n, \theta))$ имеют одинаковые точки максимума, то часто удобно использовать именно логарифмическую функцию правдоподобия. И тогда оценки можно найти, продифференцировав функцию относительно каждого оцениваемого параметра:

$$\begin{cases} \dfrac{\partial \ln L(X_1, \dots, X_n, \theta)}{\partial \theta_1} = 0 \\ \vdots \\ \dfrac{\partial \ln L(X_1, \dots, X_n, \theta)}{\partial \theta_k} = 0 \end{cases}$$

Если решить эту систему, получатся оценки максимального правдоподобия.

Приведу пример с геометрическим распределением. Пусть есть выборка $X_1, \dots, X_n$, имеющая геометрическое распределение с параметром θ – $G(\theta)$. Вероятность геометрического распределения считается так: $P(X = k) = (1-p)^{k-1}p$, где p – параметр распределения. Тогда функция правдоподобия принимает следующий вид:

$$L(X_1, \dots, X_n, \theta) = \prod_{i=1}^{n} P(X = x_i, \theta) = \prod_{i=1}^{n} (1-\theta)^{x_i - 1} \cdot \theta = \theta^n \prod_{i=1}^{n} \frac{(1-\theta)^{x_i}}{1-\theta} = \frac{\theta^n}{(1-\theta)^n} \prod_{i=1}^{n} (1-\theta)^{x_i} =$$

$= \frac{\theta^n}{(1-\theta)^n}(1-\theta)^{\sum\limits_{i=1}^{n} x_i}$

Посчитаем логарифмическую функцию правдоподобия:

$$\ln L(X_1, \dots, X_n, \theta) = \ln\left(\frac{\theta^n}{(1-\theta)^n}(1-\theta)^{\sum\limits_{i=1}^{n} x_i}\right) = n\ln\theta - n\ln(1-\theta) + \sum_{i=1}^{n} x_i \ln(1-\theta)$$

И продифференцируем для того, чтобы найти точку максимума:

$$\frac{\partial \ln L(X_1, \dots, X_n, \theta)}{\partial\theta} = \frac{n}{\theta} + \frac{n}{1-\theta} - \sum_{i=1}^{n} x_i \cdot \frac{1}{1-\theta} = 0$$

$$n + \frac{n\theta}{1-\theta} - \sum_{i=1}^{n} x_i \cdot \frac{\theta}{1-\theta} = 0$$

$$n(1-\theta) + n\theta - \sum_{i=1}^{n} x_i\theta = 0$$

$$\theta = \frac{n}{\sum\limits_{i=1}^{n} x_i} = \frac{1}{\bar{X}}$$

$$\hat{\theta} = \frac{1}{\bar{X}}$$

Таким образом, получили уже известный факт о геометрическом распределении, что его математическое ожидание равно $\frac{1}{p}$, если p – параметр.

Доверительные интервалы

Точечные оценки – это хорошо, но в большинстве случаев, если мы попытаемся посчитать вероятность, с которой оцениваемый параметр будет равен оценке, получится, что это $\frac{1}{\infty} = 0$, потому что есть бесконечное количество значений, которые может принять параметр.

Для того, чтобы все же узнать более вероятное значение, которое примет параметр, существуют ***доверительные интервалы***. Это интервал, в который оцениваемый параметр попадает с вероятностью, равной $1-\alpha$. Границами такого интервала будут случайные величины, а вероятность $1-\alpha$ называется ***уровнем доверия*** или ***доверительной вероятностью***.

$$P(T_1(X_1, \dots, X_n) < \theta < T_2(X_1, \dots, X_n)) = 1 - \alpha$$

T_1 и T_2 – это статистики – случаные функции выборки.

Существует также понятие ***центральной статистики***, которая применяется для построения доверительных интервалов. Это такая функция $G(X_1, \dots, X_n, \theta)$, которая является непрарывной и монотонной, а также ее распределение $F_G(x)$ не зависит от θ. Тогда вероятность принимает следующий вид:

$$P(g_1 < G(X_1, \dots, X_n, \theta) < g_2) = 1 - \alpha$$

И если решить это неравенство, оставляя в центре θ и убирая все остальное в края, то получится как раз доверительный интервал для θ с уровнем доверия $1-\alpha$. g_1 и g_2 – это некоторые числа, которые мы хотим пододбрать так, чтобы интервал был как можно меньше. Сейчас пойдут примеры для нормального распределения, и станет немного понятнее.

Но для начала еще упомянем теорему Фишера. По ней, если есть какая-то выборка $X_1, \dots, X_n$, все элементы которой имеют нормальное распределение $N(\mu, \sigma^2)$, то:

1. среднее выборочное $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ или $\frac{(\bar{X}-\mu)\sqrt{n}}{\sigma} \sim N(0,1)$

2. $\frac{nS^2}{\sigma^2} \sim \chi^2_{n-1}$ (S^2 – смещенная выборочная дисперсия)

3. $\bar{X}$ и S^2 независимы

4. $\frac{(\bar{X}-\mu)\sqrt{n-1}}{\sqrt{S^2}} \sim t(n-1)$

Итак, переходим к конкретным распределениям.

Доверительные интервалы для выборки из нормального распределения

1. Пусть есть выборка $X_1, \ldots, X_n \sim N(\theta_1, \sigma^2)$ с неизвестным параметром θ_1, который мы хотим оценить с помощью доверительного интервала с уровнем надежности $1-\alpha$.

Из теоремы Фишера известно, что случайная величина $\frac{(\bar{X}-\theta_1)\sqrt{n}}{\sigma} \sim N(0,1)$, а стандартное нормальное распределение не зависит от θ_1. И также известно, что функция $G(X_1, \ldots, X_n, \theta) = \frac{(\bar{X}-\theta_1)\sqrt{n}}{\sigma}$ непрерывна и монотонна. Значит, она является центральной статистикой. Тогда подберем g_1, g_2 так, чтобы доверительный интервал оказался минимальной длины:

$$P(g_1 < \frac{(\bar{X}-\theta_1)\sqrt{n}}{\sigma} < g_2) = 1-\alpha$$

Так как нормальное распределение симметрично относительно оси OY, то логично, что интервал наименьшей длины тоже должен быть симметричен. И так как есть ограничение по вероятности – она должна быть равна $1-\alpha$, то можно точно утверждать, что границами интервала наименьшей длины будут квантили стандартного нормального распределения $g_1 = z_{\alpha/2} = -z_{1-\alpha/2}$ (так как распределение симметрично) и $g_2 = z_{1-\alpha/2}$. И тогда, подставив эти квантили, получаем:

$$P(-z_{1-\alpha/2} < \frac{(\bar{X}-\theta_1)\sqrt{n}}{\sigma} < z_{1-\alpha/2}) = 1-\alpha$$
$$P(-\frac{\sigma z_{1-\alpha/2}}{\sqrt{n}} - \bar{X} < -\theta_1 < \frac{\sigma z_{1-\alpha/2}}{\sqrt{n}} - \bar{X}) = 1-\alpha$$
$$P(\bar{X} - \frac{\sigma z_{1-\alpha/2}}{\sqrt{n}} < \theta_1 < \bar{X} + \frac{\sigma z_{1-\alpha/2}}{\sqrt{n}}) = 1-\alpha$$

Таким образом, получился доверительный интервал для θ_1 : $(\bar{X} - \frac{\sigma z_{1-\alpha/2}}{\sqrt{n}}; \bar{X} + \frac{\sigma z_{1-\alpha/2}}{\sqrt{n}})$

2. Пусть есть выборка $X_1, \ldots, X_n \sim N(\mu, \theta_2^2)$ с неизвестным параметром θ_2^2, который мы хотим оценить с помощью доверительного интервала с уровнем надежности $1-\alpha$.

$X_i \sim N(\mu, \theta_2^2) \Rightarrow \frac{X_i - \mu}{\theta_2} \sim N(0,1)$. По определению распределения хи-квадрат получаем, что $\sum\limits_{i=1}^{n} \left(\frac{X_i-\mu}{\theta_2}\right)^2 \sim \chi^2(n)$. То есть распределение не зависит от θ_2, а сама функция непрерывна и монотонна. Значит, это подходящая нам центральная статистика. Так же, как и в первом случае, возьмем границами соответствующие квантили:

$$P\left(\chi^2_{n,\alpha/2} < \sum_{i=1}^{n}\left(\frac{X_i-\mu}{\theta_2}\right)^2 < \chi^2_{n,1-\alpha/2}\right) = 1-\alpha$$

$$P\left(\frac{\chi^2_{n,\alpha/2}}{\sum\limits_{i=1}^{n}(X_i-\mu)^2} < \frac{1}{\theta_2^2} < \frac{\chi^2_{n,1-\alpha/2}}{\sum\limits_{i=1}^{n}(X_i-\mu)^2}\right) = 1-\alpha$$

$$P\left(\frac{\sum\limits_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{n,1-\alpha/2}} < \theta_2^2 < \frac{\sum\limits_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{n,\alpha/2}}\right) = 1-\alpha$$

Получили доверительный интервал для θ_2^2 : $\left(\frac{\sum\limits_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{n,1-\alpha/2}}; \frac{\sum\limits_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{n,\alpha/2}}\right)$

3. Пусть есть выборка $X_1, \ldots, X_n \sim N(\theta_1, \theta_2^2)$ с неизвестным параметром θ_1 и неизвестным параметром θ_2^2, который мы хотим оценить с помощью доверительного интервала с уровнем надежности $1-\alpha$.

Воспользуемся снова теоремой Фишера. $\frac{nS^2}{\theta_2^2} \sim \chi^2_{n-1}$ – распределение не зависит от θ_2^2, функция непрерывна и монотонна, а значит можно взять как центральную статистику. Границами опять возьмем квантили:

$$P\left(\chi^2_{n-1,\alpha/2} < \frac{nS^2}{\theta_2^2} < \chi^2_{n-1,1-\alpha/2}\right) = 1-\alpha$$

$$P\left(\frac{\chi^2_{n-1,\alpha/2}}{nS^2} < \frac{1}{\theta_2^2} < \frac{\chi^2_{n-1,1-\alpha/2}}{nS^2}\right) = 1-\alpha$$

$$P\left(\frac{nS^2}{\chi^2_{n-1,1-\alpha/2}} < \theta_2^2 < \frac{nS^2}{\chi^2_{n-1,\alpha/2}}\right) = 1-\alpha$$

Получили доверительный интервал для θ_2^2 : $\left(\frac{nS^2}{\chi^2_{n-1,1-\alpha/2}}; \frac{nS^2}{\chi^2_{n-1,\alpha/2}}\right)$

4. Пусть есть выборка $X_1, \ldots, X_n \sim N(\theta_1, \theta_2^2)$ с неизвестным параметром θ_2^2 и неизвестным параметром θ_1, который мы хотим оценить с помощью доверительного интервала с уровнем надежности $1-\alpha$.

И снова воспользуемся теоремой Фишера. $\frac{(\bar{X}-\theta_1)\sqrt{n-1}}{\sqrt{S^2}} \sim t(n-1)$ – распределение не зависит от θ_1, функция непрерывна и монотонна, а значит можно взять как центральную статистику. Границами опять возьмем квантили:

$$P\left(t_{n-1,\alpha/2} < \frac{(\bar{X}-\theta_1)\sqrt{n-1}}{\sqrt{S^2}} < t_{n-1,1-\alpha/2}\right) = 1-\alpha$$

$$P\left(\frac{\sqrt{S^2}\cdot t_{n-1,\alpha/2}}{\sqrt{n-1}} - \bar{X} < -\theta_1 < \frac{\sqrt{S^2}\cdot t_{n-1,1-\alpha/2}}{\sqrt{n-1}} - \bar{X}\right) = 1-\alpha$$

$$P\left(-\frac{\sqrt{S^2}\cdot t_{n-1,1-\alpha/2}}{\sqrt{n-1}} - \bar{X} < -\theta_1 < \frac{\sqrt{S^2}\cdot t_{n-1,1-\alpha/2}}{\sqrt{n-1}} - \bar{X}\right) = 1-\alpha$$

$$P\left(\bar{X} - \frac{\sqrt{S^2}\cdot t_{n-1,1-\alpha/2}}{\sqrt{n-1}} < \theta_1 < \bar{X} + \frac{\sqrt{S^2}\cdot t_{n-1,1-\alpha/2}}{\sqrt{n-1}}\right) = 1-\alpha$$

Получили доверительный интервал для θ_1 : $\left(\bar{X} - \frac{\sqrt{S^2}\cdot t_{n-1,1-\alpha/2}}{\sqrt{n-1}}; \bar{X} + \frac{\sqrt{S^2}\cdot t_{n-1,1-\alpha/2}}{\sqrt{n-1}}\right)$

Обычно все эти большие формулы для разных доверительных интервалов сложно запомнить, поэтому может быть легче понять, как они выводятся, и запомнить только теорему Фишера и определение распределения хи-квадрат.

Статистические гипотезы

Статистическая гипотеза – это какое-то предположение о распределении случайной величины, либо о параметрах распределения, либо о свойствах распределения. Мы выдвигаем некоторое предположение, которое не хотим отвергать без явных свидетельств его ошибочности, это предположение называется ***нулевой гипотезой*** (обозначается H_0), и она как раз-таки будет проверяться. В противовес ей идет ***альтернативная гипотеза*** (обозначается $H_A, H_1, H_2, \ldots$) – это отклонение от основной гипотезы, которое нам важно выявить (если оно действительно есть).

Чтобы принять решение, какую гипотезу выбрать в результате проверки, нужно еще до начала проверки установить ***статистический критерий*** – правило, на основании которого принимается решение, отвергнуть основную гипотезу в пользу альтернативной или нет, в зависимости от результатов наблюдений. Это правило устанавливается так: берется функция от выборки $T(X_1, \ldots, X_n)$, для которой известно распределение при верной основной гипотезе. Такая функция называется ***статистикой критерия***. После этого определяются границы ***критической области*** (или областей, если их несколько) – это область значений статистики, при попадании в которую будет отвергаться основная гипотеза. Оставшаяся область называется ***доверительной областью*** – при попадании в нее значения статистики основная гипотеза не будет отвергнута.

Также до начала проверки задается ***уровень значимости*** – допустимая вероятность отвергнуть основную гипотезу, когда она верна. Обычно уровень значимости берут $\alpha = 0{,}05$.

Итак, как же проверять гипотезу? Общий алгоритм действий выглядит следующим образом:
1. Формулируем основную и альтернативную гипотезы.
2. Выбираем уровень значимости.
3. Выбираем статистику.
4. Определяем, какое будет распределение у статистики, если верна основная гипотеза.
5. Строим доверительную и критическую области (определяем их границы).
6. Вычисляем реализацию статистики от нашей выборки.
7. Смотрим, куда попало значение статистики и принимаем решение, отвергнуть нулевую гипотезу или нет.

Существует несколько известных критериев для проверки гипотез о разных распределениях. Рассмотрим их все.

Биномиальный критерий

Есть выборка $X_1, \ldots, X_n \sim Bi(1, p)$, где p неизвестно, и мы хотим выдвинуть гипотезу о его значении. Пойдем по пунктам алгоритма:
1. $H_0 : p = p_0$ (p_0 – это какое-то конкретное значение, например 0,5)
   $H_1 : p > p_0$
2. Выбираем какое-то значение α (допустим, 0,05)
3. $T(X_1, \ldots, X_n) = \sum\limits_{i=1}^{n} X_i$ (это равно числу успехов)
4. $T(X_1, \ldots, X_n)|_{H_0} \sim Bi(n, p_0)$
5. Доверительная область будет в границах $(-\infty; z_{1-\alpha}]$, а критическая область, соответственно, в границах $(z_{1-\alpha}; +\infty)$ ($z_{1-\alpha}$ – это квантиль распределения $Bi(n, p_0)$)

Остальные два пункта выполняются уже на конкретных выборках. По формуле из 3 пункта считается значение статистики, потом надо посмотреть, в какой промежуток попало число. Если в доверительный интервал, то говорят, что нулевая гипотеза не отвергается. Если в критический

интервал, то говорят, что нулевая гипотеза отвергается в пользу альтернативной.

Гипотезы в гауссовских моделях

1. Выдвигаем гипотезу о значении математического ожидании при известной дисперсии. Выборка $X_1, \ldots, X_n \sim N(m, \sigma^2)$

1. $H_0 : m = m_0$ (m_0 – это какое-то конкретное значение, например 1)
   $H_1 : m < m_0$
   $H_2 : m > m_0$
   $H_3 : m \neq m_0$

Здесь это записано так для удобства, чтобы показать сразу, как проверяются несколько альтернативных гипотез. В реальных проверках гипотез всегда выдвигает одна альтернативная гипотеза, и уже относительно нее проводятся дальнейшие вычисления. Здесь проверка гипотез будет отличаться только границами доверительной и критической областей.

2. некоторое α

3. $T(X_1, \ldots, X_n) = \dfrac{(\bar{X} - m_0)\sqrt{n}}{\sigma}$

4. $T(X_1, \ldots, X_n)|_{H_0} \sim N(0, 1)$

5. Для H_1 доверительная область в границах $[z_\alpha; +\infty)$, критическая область в границах $(-\infty; z_\alpha)$
Для H_2 доверительная область в границах $(-\infty; z_{1-\alpha}]$, критическая область в границах $(z_{1-\alpha}; +\infty)$
Для H_3 доверительная область в границах $[z_{\alpha/2}; z_{1-\alpha/2}]$, критическая область в границах $(-\infty; z_{\alpha/2}) \cup (z_{1-\alpha/2}; +\infty)$

2. Выдвигаем гипотезу о значении математического ожидании при неизвестной дисперсии. Выборка $X_1, \ldots, X_n \sim N(m, \sigma^2)$

1. $H_0 : m = m_0$
   $H_1 : m < m_0$
   $H_2 : m > m_0$
   $H_3 : m \neq m_0$

2. некоторое α

3. $T(X_1, \ldots, X_n) = \dfrac{(\bar{X} - m_0)\sqrt{n}}{\tilde{S}}$ ($\tilde{S}$ – корень из несмещенной выборочной лисперсии)

4. $T(X_1, \ldots, X_n)|_{H_0} \sim t(n-1)$

5. Для H_1 доверительная область в границах $[t_{\alpha,n-1}; +\infty)$, критическая область в границах $(-\infty; t_{\alpha,n-1})$
Для H_2 доверительная область в границах $(-\infty; t_{1-\alpha,n-1}]$, критическая область в границах $(t_{1-\alpha,n-1}; +\infty)$
Для H_3 доверительная область в границах $[t_{\alpha/2,n-1}; t_{1-\alpha/2,n-1}]$, критическая область в границах $(-\infty; t_{\alpha/2,n-1}) \cup (t_{1-\alpha/2,n-1}; +\infty)$

3. Выдвигаем гипотезу о значении дисперсии при известном математическом ожидании. Выборка $X_1, \ldots, X_n \sim N(m, \sigma^2)$

1. $H_0 : \sigma^2 = \sigma_0^2$
   $H_1 : \sigma^2 \neq \sigma_0^2$

2. некоторое α

3. $T(X_1, \ldots, X_n) = \sum\limits_{i=1}^{n} \dfrac{(X_i - m)^2}{\sigma_0^2}$

4. $T(X_1, \ldots, X_n)|_{H_0} \sim \chi^2(n)$

5. Доверительная область в границах $[\chi^2_{\alpha/2,n}; \chi^2_{1-\alpha/2,n}]$, критическая область в границах $(-\infty; \chi^2_{\alpha/2,n}) \cup (\chi^2_{1-\alpha/2,n}; +\infty)$

4. Выдвигаем гипотезу о значении дисперсии при неизвестном математическом ожидании. Выборка $X_1, \ldots, X_n \sim N(m, \sigma^2)$

1. $H_0 : \sigma^2 = \sigma_0^2$
   $H_1 : \sigma^2 \neq \sigma_0^2$
2. некоторое α
3. $T(X_1, \ldots, X_n) = \sum\limits_{i=1}^{n} \dfrac{(X_i - \bar{X})^2}{\sigma_0^2}$
4. $T(X_1, \ldots, X_n)|_{H_0} \sim \chi^2(n-1)$
5. Доверительная область в границах $[\chi^2_{\alpha/2,n-1}; \chi^2_{1-\alpha/2,n-1}]$, критическая область в границах $(-\infty; \chi^2_{\alpha/2,n-1}) \cup (\chi^2_{1-\alpha/2,n-1}; +\infty)$

Критерий Стьюдента

Пусть есть две выборки:
$X_1, \ldots, X_{n_1} \sim N(m_1, \sigma_1^2)$
$Y_1, \ldots, Y_{n_2} \sim N(m_2, \sigma_2^2)$
В них неизвестны все параметры, но $\sigma_1 = \sigma_2$ и все случайные величины независимы.
1. $H_0 : m_1 - m_2 = 0$
   $H_1 : m_1 - m_2 < 0$
   $H_2 : m_1 - m_2 > 0$
   $H_3 : m_1 - m_2 \neq 0$
2. некоторое α
3. $T(X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2}) = \dfrac{\bar{X} - \bar{Y}}{\hat{\sigma}_1 \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$

$$\hat{\sigma}_1^2 = \frac{\sum\limits_{i=1}^{n_1}(X_i - \bar{X})^2 + \sum\limits_{i=1}^{n_2}(Y_i - \bar{Y})^2}{n_1 + n_2 - 2}$$

4. $T(X_1, \ldots, X_n)|_{H_0} \sim t(n_1 + n_2 - 2)$
5. Для H_1 доверительная область в границах $[t_{\alpha,n_1+n_2-2}; +\infty)$, критическая область в границах $(-\infty; t_{\alpha,n_1+n_2-2})$
Для H_2 доверительная область в границах $(-\infty; t_{1-\alpha,n_1+n_2-2}]$, критическая область в границах $(t_{1-\alpha,n_1+n_2-2}; +\infty)$
Для H_3 доверительная область в границах $[t_{\alpha/2,n_1+n_2-2}; t_{1-\alpha/2,n_1+n_2-2}]$, критическая область в границах
$(-\infty; t_{\alpha/2,n_1+n_2-2}) \cup (t_{1-\alpha/2,n_1+n_2-2}; +\infty)$

Критерий Фишера

Пусть есть две выборки:
$X_1, \ldots, X_{n_1} \sim N(m_1, \sigma_1^2)$
$Y_1, \ldots, Y_{n_2} \sim N(m_2, \sigma_2^2)$
В них неизвестны все параметры, но все случайные величины независимы.
1. $H_0 : \sigma_1^2 = \sigma_2^2$
   $H_1 : \sigma_1^2 \neq \sigma_2^2$
2. некоторое α
3. $T(X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2}) = \dfrac{\tilde{S}_X^2}{\tilde{S}_Y^2}$
4. $T(X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2})|_{H_0} \sim F(n_1 - 1, n_2 - 1)$
5. Доверительная область в границах $[F_{\alpha/2,n_1-1,n_2-1}; F_{1-\alpha/2,n_1-1,n_2-1}]$, критическая область в границах $[0; F_{\alpha/2,n_1-1,n_2-1}) \cup (F_{1-\alpha/2,n_1-1,n_2-1}; +\infty)$

Критерий проверки некоррелированности двух случайных величин

Пусть есть выборка из пар случайных величин $(X_1, Y_1), \ldots, (X_n, Y_n)$, порожденная гауссовским случайным вектором (X, Y)

1. $H_0 : \rho_{xy} = 0$
   $H_1 : \rho_{xy} < 0$
   $H_2 : \rho_{xy} > 0$
   $H_3 : \rho_{xy} \neq 0$
2. некоторое α
3. $T((X_1, Y_1), \ldots, (X_n, Y_n)) = \dfrac{\sqrt{n-2}\hat{\rho}_{xy}}{\sqrt{1-\hat{\rho}_{xy}^2}}$

$$\hat{\rho}_{xy} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

4. $T((X_1, Y_1), \ldots, (X_n, Y_n))|_{H_0} \sim t(n-2)$
5. Для H_1 доверительная область в границах $[t_{\alpha,n-2}; +\infty)$, критическая область в границах $(-\infty; t_{\alpha,n-2})$
   Для H_2 доверительная область в границах $(-\infty; t_{1-\alpha,n-2}]$, критическая область в границах $(t_{1-\alpha,n-2}; +\infty)$
   Для H_3 доверительная область в границах $[t_{\alpha/2,n-2}; t_{1-\alpha/2,n-2}]$, критическая область в границах $(-\infty; t_{\alpha/2,n-2}) \cup (t_{1-\alpha/2,n-2}; +\infty)$

Существуют еще два критерия, но они посложнее, и на экзамене не попадутся. Приведу еще пример решения задачи на доверительный интервал и на проверку гипотезы из нулевика прошлого года, чтобы было понятно, как это все используется, хотя в целом, по сути просто надо подобрать формулу под нужную ситуацию, подставить числа и аккуратно все посчитать.

Задание: Из произведённой партии шоколадных батончиков случайным образом были выбраны шесть штук, вес которых (в граммах) составил 49.1; 50,3; 49.6; 51,2; 48.4; 49.8. Постройте центральный доверительный интервал уровня надёжности 0,95 для среднего веса батончика. Проверьте гипотезу о том, что средний вес батончика составляет 50 г. Уровень значимости принять равным 0.1. Предполагается, что наблюдения имеют гауссовское распределение.

Решение:

Вначале проведем вспомогательные вычисления, то есть посчитаем выборочное математическое ожидание и несмещенную выборочную дисперсию:

$$\overline{X} = \frac{1}{6}\sum_{i=1}^{6} X_i = \frac{1}{6}(49{,}1 + 50{,}3 + 49{,}6 + 51{,}2 + 48{,}4 + 49{,}8) = \frac{298{,}4}{6} = 49{,}73$$

$$\tilde{S}^2 = \frac{1}{5}\sum_{i=1}^{6}(X_i - 49{,}73)^2 = \frac{1}{5}(0{,}3969 + 0{,}3249 + 0{,}0169 + 2{,}1609 + 1{,}7689 + 0{,}0049) = \frac{4{,}6734}{5} = 0{,}9347$$

$$\tilde{S} = \sqrt{0{,}9347} = 0{,}9668$$

Так как неизвестно ни математическое ожидание рассматриваемого распределения, ни дисперсия, то доверительный интервал уровня надежности 0,95 будет иметь следующий вид:

$$P(\overline{X} - \frac{\tilde{S}t_{5;0{,}975}}{\sqrt{6}} < \theta_1 < \overline{X} + \frac{\tilde{S}t_{5;0{,}975}}{\sqrt{6}}) = 0{,}95$$

$$P(49{,}73 - \frac{0{,}9668 \cdot 2{,}5706}{2{,}4495} < \theta_1 < 49{,}73 + \frac{0{,}9668 \cdot 2{,}5706}{2{,}4495}) = 0{,}95$$

$$P(48{,}7154 < \theta_1 < 50{,}7446) = 0{,}95$$

Доверительный интервал уровня надежности 0,95 равен $(48{,}7154; 50{,}7446)$

Проверим гипотезу о среднем весе батончика. Выборка распределена по гауссовскому закону, но ни математическое ожидание, ни дисперсия неизвестны.
$X_1, \dots, X_6 \sim N(m, \sigma^2)$
Пройдем по всем пунктам проверки гипотезы.
1. Основная гипотеза: $H_0 : m = 50$, альтернативная гипотеза: $H_1 : m \neq 50$
2. $\alpha = 0{,}1$
3. $T(X) = \dfrac{(\overline{X} - 50)\sqrt{6}}{\tilde{S}} = \dfrac{(49{,}73 - 50)\sqrt{6}}{0{,}9668} = -\dfrac{0{,}6614}{0{,}9668} = -0{,}684$
4. $T(X)|_{H_0} \sim t(6)$
5. Доверительная область лежит в границах от $-t_{5;0,95}$ до $t_{5;0,95}$, то есть от $-2{,}015$ до $2{,}015$. Полученное в пункте 3 значение попадает в эту область, значит, гипотеза верна на уровне значимости 0,1.

Ответ: Доверительный интервал: $(48{,}7154; 50{,}7446)$; гипотеза верна.