Пусть теперь ξ – СВНТ. Не известна плотность ее распределения $f_\xi(x)$. Для оценки плотности распределения по выборке $\xi_1, \xi_2, \ldots, \xi_n$ сделаем следующее:

- разобьем область значений случайной величины ξ на интервалы длины h_i $(i = 1, 2, \ldots, s)$;
- найдем x_i^* – середины i-го интервала;
- подсчитаем число ν_i – число элементов в выборке, попавших в i-ый интервал;
- вычислим значение $f_n(x_i^*) = \dfrac{\nu_i}{nh_i}$, являющееся оценкой плотности распределения случайной величины в точке x_i^*.

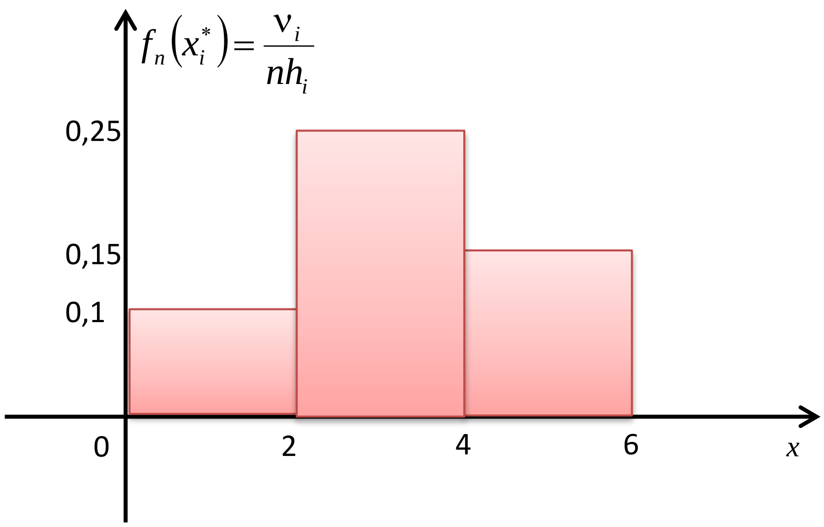
Определение. *Фигура, составленная из прямоугольников с основаниями* h_i $(i = 1, 2, \ldots, s)$ *и высотами* $\dfrac{\nu_i}{nh_i}$, называется **гистограммой.**

Определение. *Эмпирической функцией распределения назовем функцию распределения СВДТ, принимающей значения середин выбранных интервалов (* x_i^* *) с вероятностям, равными* $\dfrac{\nu_i}{n}$.

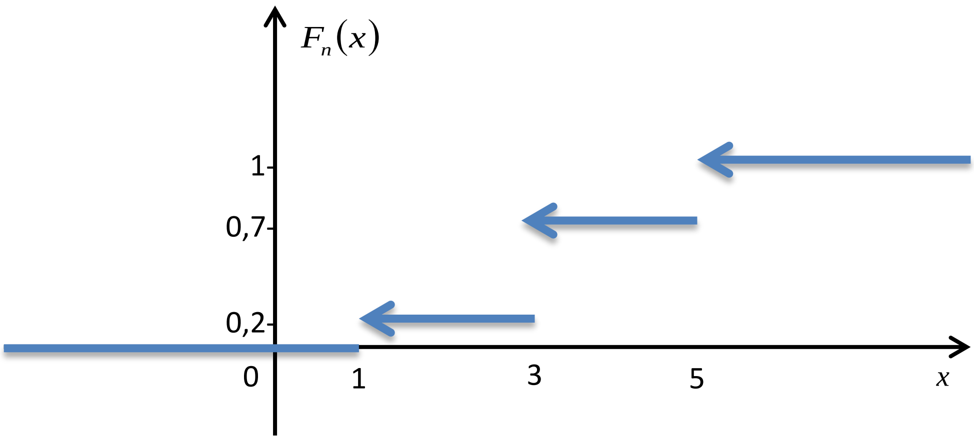
Пример. Пусть имеется выборка объема 100 из генеральной совокупности с теоретической плотностью распределения $f_\xi(x)$. Результаты измерений занесены в таблицу:

№ интервала i	Границы интервала (x_i, x_{i+1})	Середина интервала x_i^*	Число элементов, попавших в интервал ν_i	$f_n(x_i^*) = \dfrac{\nu_i}{nh_i}$
1	(0, 2)	1	20	0,1
2	(2, 4)	3	50	0,25
3	(4, 6)	5	30	0,15

Гистограмма



Эмпирическая функция распределения



§2. Точечные оценки параметров распределений.

Смысл точечной оценки: исследуемая величина считается приближенно равной вычисляемому значению.

Определение. Пусть $\xi_1, \xi_2, \ldots, \xi_n$ – наблюдаемые значения случайной величины ξ.

- $\bar{\xi} = \frac{1}{n}\sum_{i=1}^{n}\xi_i$ – выборочное среднее случайной величины ξ.
- $s^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\xi_i - \bar{\xi}\right)^2$ – выборочная дисперсия.
- $s_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\xi_i - \bar{\xi}\right)^2$ – исправленная выборочная дисперсия.

Замечание. По своей сути, точечные оценки случайной величины сами являются случайными величинами, но в статистических расчетах вместо случайных величин подставляют их числовые значения. Для выборочного среднего принято обозначение $\bar{x}_{\text{в}}$.

Типы точечных оценок.

Пусть ξ_i – случайная величина, представляющая результат i-го испытания величины ξ, a – оцениваемый параметр, $\tilde{a}$ – его точечная оценка. Тогда эта оценка является функцией от случайных величин $\xi_1, \xi_2, \ldots, \xi_n$, то есть $\tilde{a} = \varphi\left(\xi_1, \xi_2, \ldots, \xi_n\right)$.

Определение. *Точечная оценка* $\tilde{a}$ *называется* ***состоятельной****, если* $\forall \varepsilon > 0 \quad \lim\limits_{n\to\infty} P\left(\left|a - \varphi\left(\xi_1, \xi_2, \ldots, \xi_n\right)\right| < \varepsilon\right) = 1$, *в этом случае говорят, что точечная оценка* $\tilde{a}$ *сходится к оцениваемому параметру по вероятности.*

Определение. *Точечная оценка* $\tilde{a}$ *называется* ***несмещенной****, если ее математическое ожидание равно оцениваемому параметру, то есть* $\mathbf{M}\tilde{a} = a$.

Определение. *Точечная оценка* $\tilde{a}$ *называется* ***эффективной****, если ее дисперсия минимальна.*

Справедливая следующая теорема.

Теорема 1. Выборочное среднее $\overline{\xi} = \frac{1}{n}\sum_{i=1}^{n}\xi_i$ – состоятельная несмещенная оценка математического ожидания. В случае, когда случайная величина ξ подчинена нормальному закону распределения, то эта оценка является также эффективной.

Теорема 2. Выборочная дисперсия $s^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\xi_i - \overline{\xi}\right)^2$ – состоятельная оценка дисперсии $\mathbf{D}\xi$, которая, однако, не является несмещенной. Исправленная выборочная дисперсия $s_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\xi_i - \overline{\xi}\right)^2$ является состоятельной несмещенной оценкой дисперсии.

Замечание. Легко заметить, что $s_1^2 = \frac{n}{n-1}s^2$, $\frac{n}{n-1}\underset{n\to\infty}{\to}1$, и при больших объемах выборки $s_1^2 \approx s_1$. При малых значениях *n* лучше брать исправленную дисперсию s_1^2.

§3. Интервальные оценки параметров распределений.

Смысл интервальной оценки: исследуемая величина с заданной вероятностью попадает в данный интервал.

Пусть ξ – случайная величина, подчиняющаяся нормальному закону $N(a,\sigma)$. Неизвестная плотность распределения имеет вид: $f_\xi(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-a)^2}{2\sigma^2}}$, при этом неизвестны именно параметры распределения. Пусть ξ_i – случайная величина, представляющая результат *i*-го испытания величины ξ, тогда $\mathbf{M}\xi_i = a$, $\mathbf{D}\xi_i = \sigma^2$. Можно доказать, что выборочное среднее $\overline{\xi} = \frac{1}{n}\sum_{i=1}^{n}\xi_i$ также является нормально распределенной случайной величиной с числовыми характеристиками $\mathbf{M}\overline{\xi} = a$, $\mathbf{D}\overline{\xi} = \frac{\sigma^2}{n}$. То есть выборочное среднее подчиняется закону $N\left(a,\frac{\sigma}{\sqrt{n}}\right)$.

Рассмотрим вероятность того, что выборочное среднее, как нормальная случайная величина отклоняется от своего математического ожидания на некоторую величину $\delta > 0$.

$$P\left(\left|\overline{\xi} - a\right| < \delta\right) = P\left(a - \delta < \overline{\xi} < a + \delta\right) = \Phi\left(\frac{a+\delta-a}{\sqrt{\mathbf{D}\overline{\xi}}}\right) - \Phi\left(\frac{a-\delta-a}{\sqrt{\mathbf{D}\overline{\xi}}}\right) = 2\Phi\left(\frac{\delta\sqrt{n}}{\sigma}\right). \quad (*)$$

Обозначим $t = \frac{\delta\sqrt{n}}{\sigma}$, тогда $P\left(\left|\overline{\xi} - a\right| < \delta\right) = 2\Phi(t)$.

Вероятность $2\Phi(t)$ называют доверительной вероятностью. Обычно считают, что $2\Phi(t)\approx 1$, или $2\Phi(t)=1-\alpha$, где $\alpha-$достаточно маленькое положительное число, называемое уровнем значимости.

Доверительный интервал – интервал, который с заданной доверительной вероятностью покрывает оцениваемый параметр.

Пусть известен параметр σ, а искомым параметром является a. $t=\dfrac{\delta\sqrt{n}}{\sigma}\Rightarrow\delta=\dfrac{\sigma t}{\sqrt{n}}$, а тогда, если раскрыть модуль в выражении (*) по-другому, получим:

$$P\left(\overline{\xi}-\delta<a<\overline{\xi}+\delta\right)=P\left(a\in\left(\overline{\xi}-\delta,\overline{\xi}+\delta\right)\right)=2\Phi(t)=1-\alpha.$$

Интервал $\left(\overline{\xi}-\delta,\overline{\xi}+\delta\right)$ является доверительным интервалом. Для нахождения числа δ нужно по таблице функции Лапласа найти такое значение t, для которого $\Phi(t)=\dfrac{1-\alpha}{2}$.

Замечание. Если параметр σ также не известен, то можно заменить, используя исправленную дисперсию, то есть $\sigma\approx\sqrt{s_1^2}$.

§4. Статистическая проверка гипотезы о характере распределения. Критерий Пирсона (критерий χ^2).

Пусть $\xi_1,\xi_2,\ldots,\xi_n$ – выборка из генеральной совокупности с теоретической функцией распределения $F_\xi(x)$. Выдвигаем гипотезу H_0: выборка $\xi_1,\xi_2,\ldots,\xi_n$ соответствует случайной величине с теоретической функцией распределения $F_0(x)$. $F_0(x)$ – некоторая конкретная функция. Для проверки гипотезы H_0 используем критерий Пирсона.

1. Найдем размах выборки и выберем подходящий отрезок, в котором принимает значения случайная величина. Этот отрезок разобьем на r интервалов $\Delta_1,\Delta_2,\ldots,\Delta_r$, $\Delta_i=\left(a_i,a_{i+1}\right)$.
2. Найдем вероятность попадания случайной величины, распределенной по закону $F_0(x)$ в каждый из этих интервалов. $p_i=P\left(\xi\in\Delta_i\right)=F_0(a_{i+1})-F_0(a_i)=\int\limits_{a_i}^{a_{i+1}}f_0(x)dx$.
3. Пусть ν_i – число элементов в выборке, которые попали в Δ_i (эмпирическая частота). Число значений случайной величины, попавшей в интервал Δ_i – случайная величина, подчиняющаяся биномиальному закону с параметрами $n-$ число испытаний (число измерений случайной величины) и p_i-вероятность успеха (вероятность попадания в нужный интервал). Математическое ожидание случайной величины, распределенной по данному биномиальному закону, равно np_i. Это число np_i суть ожидаемая (теоретическая) частота.

4. За меру отклонения распределения выборки от гипотетического принимается величина $\chi^2_{\text{в}}$ («хи квадрат выборочное»), $\chi^2_{\text{в}} = \sum_{i=1}^{r} \frac{(\nu_i - np_i)^2}{np_i}$. Пирсон доказал, что в случае справедливости гипотезы H_0 распределение случайной величины $\chi^2_{\text{в}}$ при $n \to \infty$ сходится к известному распределению χ^2 ((«хи квадрат») с $r-1$ степенями свободы.
5. Зададим уровень значимости α. Обычно $\alpha = 0{,}05$. По таблице распределений χ^2 с $r-1$ степенями свободы находим такое значение $\chi^2_{\alpha,r-1}$, что $P\left(\chi^2 > \chi^2_{\alpha,r-1}\right) = \alpha$.
6. По имеющейся выборке вычисляем значение $\chi^2_{\text{в}}$. Если окажется, что $\chi^2_{\text{в}} > \chi^2_{\alpha,r-1}$, то такое отклонение считается значимым, и мы с уровнем значимости α отвергаем гипотезу H_0, то есть с вероятностью 95% эта гипотеза не является верной. Если же $\chi^2_{\text{в}} \le \chi^2_{\alpha,r-1}$, то гипотеза принимается на уровне значимости α.

Важные замечания.

- В приложениях интервалы Δ_i выбираются так, чтобы число элементов в каждом из них было не очень малым, то есть чтобы теоретическая частота $np_i \ge 5$.
- Интервалы Δ_i не обязательно должны быть одинаковой длины.
- Если распределение зависит от неизвестных параметров, то вероятности p_i попадания в интервалы Δ_i вычисляем, заменяя эти параметры достаточно хорошими точечными оценками. В этом случае число степеней свободы будет равно $r-l-1$, где $r-$ число интервалов, оставшихся после объединения, $l-$ число неизвестных параметров распределения.

Далее вы можете увидеть статистическую проверку гипотезы о распределении баллов за первый коллоквиум по теории вероятностей студентов 2 курса (2012 год).