The Chi-Squared Distribution

Let $x \sim N(0,1)$. Consider $y = x^2$. What is the density of y?

$f(y) = ?$

(why?) We consider positive quantities, so we often square quantities, but things tend to be Normal by CLT.

One approach to figure out $f(y)$ is by a change of variable

$$\int_{-\infty}^{+\infty} f(x)\,dx = 1$$

let $y = x^2 \Rightarrow x = \begin{cases} \sqrt{y} & x \geqslant 0 \\ -\sqrt{y} & x \leqslant 0 \end{cases}$

$$\frac{dy}{dx} = 2x \Rightarrow dx = \frac{dy}{2x}$$

$$\int_{-\infty}^{+\infty} f(x)\,dx = \int_{-\infty}^{0} f(x)\,dx + \int_{0}^{+\infty} f(x)\,dx$$

$$= \int_{\infty}^{0} f(-\sqrt{y})\,\frac{dy}{-2\sqrt{y}} + \int_{0}^{\infty} f(\sqrt{y})\,\frac{dy}{2\sqrt{y}}$$

Since $f(x)$ is symmetric, $f(-\sqrt{y}) = f(\sqrt{y})$

$$= \int_{0}^{\infty} f(\sqrt{y})\,\frac{dy}{2\sqrt{y}} + \int_{0}^{\infty} f(\sqrt{y})\,\frac{dy}{2\sqrt{y}} = \int_{0}^{\infty} \frac{f(\sqrt{y})}{\sqrt{y}}\,dy = 1$$

$$f(y) = \frac{f(\sqrt{y})}{\sqrt{y}} = \frac{1}{\sqrt{2\pi y}}\, e^{-y/2} \qquad y \geq 0$$

In general, if $y = g(x)$ and $x = g^{-1}(y)$ [invertible]

then $$f(y) = \frac{f_X(g^{-1}(y))}{|g'(g^{-1}(y))|}$$ for appropriate range of y

Example: $X \sim \text{Unif}(0,1)$ and $y = \ln x = g(x)$

What is $f(y)$?

$$x = e^{y} = g^{-1}(y) \quad \text{and} \quad f_X(g^{-1}(y)) = 1 \qquad y \leq 0$$

$$g'(x) = \frac{1}{x} \Rightarrow g'(g^{-1}(y)) = e^{-y}$$

$$f(y) = \frac{1}{e^{-y}} = e^{y} \qquad y \leq 0$$

Check: $$\int_{-\infty}^{0} e^{y}\,dy = e^{y}\Big|_{-\infty}^{0} = 1 - e^{-\infty} = 1.$$

Example: $x \sim \text{Unif}(0,1)$ $\quad y = -2x = g(x)$

$$x = -\frac{y}{2} = g^{-1}(y)$$

$$g'(x) = -2 \implies g'(g^{-1}(y)) = -2$$

$$f(y) = \frac{1}{|-2|} = \frac{1}{2} \qquad -2 \leqslant y \leqslant 0$$

Check: $\displaystyle\int_{-2}^{0} \frac{1}{2}\,dy = \frac{1}{2}y\Big|_{-2}^{0} = 0 + 1 = 1.$

If $X \sim N(0,1)$, then $y = x^2$ has density

$$f(y) = \frac{1}{2\sqrt{\pi}} \left(\frac{y}{2}\right)^{\frac{1}{2}-1} e^{-y/2}$$

It turns out this form can be generalized for any $k \in \mathbb{N}$.

$$f(y) = \frac{1}{2\Gamma(\frac{k}{2})} \left(\frac{y}{2}\right)^{\frac{k}{2}-1} e^{-y/2} \qquad y \geq 0$$

where $\Gamma(x)$ is the Gamma function $\int_0^\infty t^{x-1} e^{-t} \, dt$

Check: $\displaystyle\int_0^\infty \frac{1}{2\Gamma(\frac{k}{2})} \left(\frac{y}{2}\right)^{\frac{k}{2}-1} e^{-y/2} \, dy = \frac{1}{2\Gamma(\frac{k}{2})} \, 2 \underbrace{\int_0^\infty t^{\frac{k}{2}-1} e^{-t} \, dt}_{\Gamma(\frac{k}{2})} = 1$

where $t = \frac{y}{2}$

- We can show that the Gamma function $\Gamma(x)$ satisfies:

$$\Gamma(x) = (x-1)\Gamma(x-1) \qquad \text{where } x > 1.$$

$$\Gamma(x) = \int_0^\infty \underbrace{t^{x-1}}_{u}\,\underbrace{e^{-t}dt}_{dv} = \underbrace{-t^{x-1}e^{-t}}_{uv}\Big|_0^\infty - \int_0^\infty \underbrace{-e^{-t}}_{v}\,\underbrace{(x-1)\,t^{x-2}\,dt}_{du}$$

$$= 0 + (x-1)\,\Gamma(x-1)$$

- Also, $\Gamma(1) = 1$. What happens at integer values?

$$\Gamma(2) = 1\,\Gamma(1) = 1 = 1!$$

$$\Gamma(3) = 2\,\Gamma(2) = 2 = 2!$$

$$\Gamma(4) = 3\,\Gamma(3) = 6 = 3!$$

$$\vdots$$

- Also for any $a > 0$: $\quad a(a+1)(a+2)\ldots(a+k-1) = \dfrac{\Gamma(a+k)}{\Gamma(a)}$
- $\Gamma(1/2) = \sqrt{\pi}$

The Chi-Squared density

$$f(y) = \frac{1}{2\Gamma\left(\frac{k}{2}\right)} \left(\frac{y}{2}\right)^{\frac{k}{2}-1} e^{-y/2}$$

$$= \frac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)} y^{\frac{k}{2}-1} e^{-y/2} \qquad k \in \mathbb{N}$$

k is the "degree" of the distribution. $y \sim \chi^2_k$

$E[y] = k \qquad Var(y) = 2k$

- If $X \sim N(0,1) \Rightarrow x^2 \sim \chi^2_1$ (with $k=1$)
- Sum of independent χ^2 is χ^2 with degree equal to sum of degrees.

Classical applications of χ^2

Imagine we roll a pair of dice $n = 144$ times and we obtain the following outcomes for $s \in \{2, 3, \dots, 12\}$

s:	2	3	4	5	6	7	8	9	10	11	12
y_s:	2	4	10	12	22	29	21	15	14	9	6

\# times we have seen outcome s

The corresponding probabilities are:

s:	2	3	4	5	6	7	8	9	10	11	12
p_s:	$\frac{1}{36}$	$\frac{1}{18}$	$\frac{1}{12}$	$\frac{1}{9}$	$\frac{5}{36}$	$\frac{1}{6}$	$\frac{5}{36}$	$\frac{1}{9}$	$\frac{1}{12}$	$\frac{1}{18}$	$\frac{1}{36}$

Each y_s is a binomial R.V. with $n = 144$ trials and success probability p_s. So $E[y_s] = n p_s$

s:	2	3	4	5	6	7	8	9	10	11	12
y_s:	2	4	10	12	22	29	21	15	14	9	6
np_s:	4	8	12	16	20	24	20	16	12	8	4

One important question is whether the dice are fair, in other words, are the outcomes coming from the claimed distribution?

Construct $Z_s = \dfrac{y_s - np_s}{\sqrt{np_s(1-p_s)}} \sim N(0,1)$ [CLT]

- $y_s - np_s$ measures deviation from "Expected"
- Dividing by $\sqrt{np_s(1-p_s)}$ scales this, so all s are treated "equally"
- Some are positive, some are negative $\Rightarrow$ Square them!

$$\sum_{s=2}^{12} Z_s^2 \sim \chi^2_{11} \quad ? \quad \textbf{NO}$$

they are NOT independent! For instance, given

$y_2, y_3, \ldots, y_{11}$, we can determine

$$y_{12} = n - (y_2 + y_3 + \cdots + y_{11})$$

It turns out, theoretically, the correct thing to do is

$$\sum_s Z_s^{*2} = \sum_s \left(\frac{y_s - np_s}{\sqrt{np_s}}\right)^2 \sim \chi^2_{10}$$

χ^2 test

In general, $\displaystyle\sum_s \frac{(\# - \text{Expected})^2}{\text{Expected}} \sim \chi^2_{K-1}$

Check P-value using χ^2_{K-1}

Proof for $k=2$

$$Z_1^{*2} + Z_2^{*2} = \frac{(y_1 - np_1)^2}{np_1} + \frac{(y_2 - np_2)^2}{np_2}$$

use $y_2 = n - y_1$

we get

$$\frac{(y_1 - np_1)^2}{np_1} + \frac{(-y_1 + np_1)^2}{n(1-p_1)}$$

$$= \frac{(y_1 - np_1)^2}{np_1(1-p_1)} = Z_1^2 \sim \chi_1^2 = \chi_{k-1}^2$$

Typical χ^2 test examples

Contingency tables

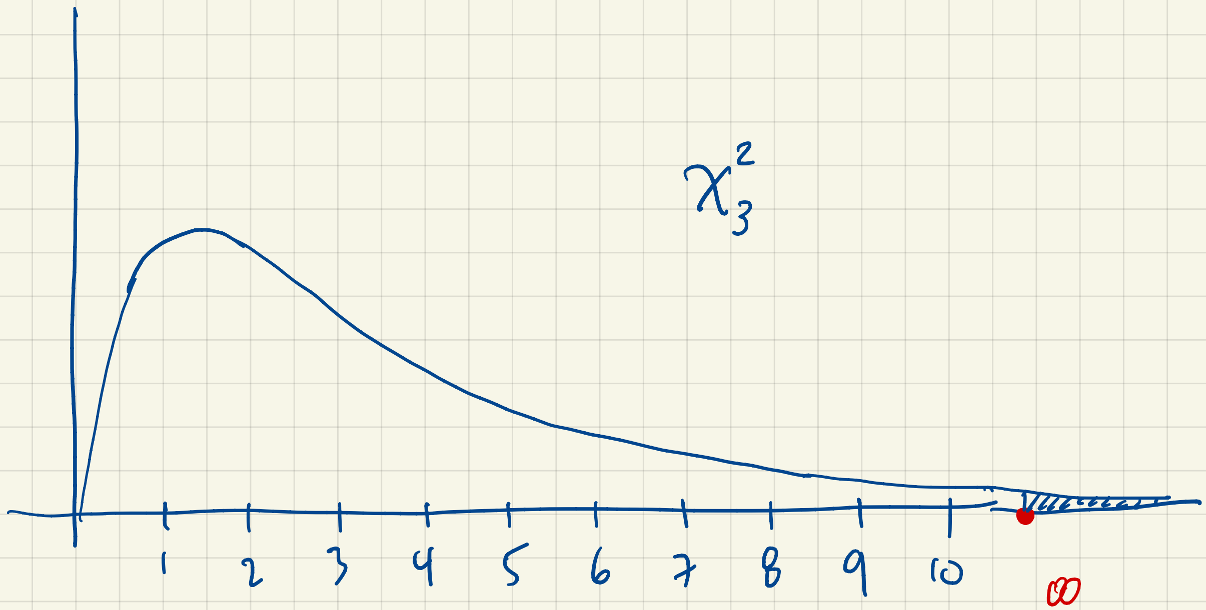
	A	B
A	14	6
B	16	24

Check if outcomes come from a uniform dist $\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)$

$n = 60.$ $\quad np_s = \frac{60}{4} = 15.$

$$\frac{(14-15)^2}{15} + \frac{(6-15)^2}{15} + \frac{(16-15)^2}{15} + \frac{(24-15)^2}{15} = 10.933$$

check Pvalue on χ^2_3 $\quad (k-1=3)$



$$\int_{10.933}^{\infty} f(y)\,dy \quad \text{where } y \sim \chi^2_3$$

What if we want to check for <u>independence</u> of traits.

	$\frac{1}{2}$	$\frac{1}{2}$		
	A	B		
A	14	6	20	$\frac{1}{3}$
B	16	24	40	$\frac{2}{3}$
	30	30	60	

Are and independent?

$$\Downarrow$$

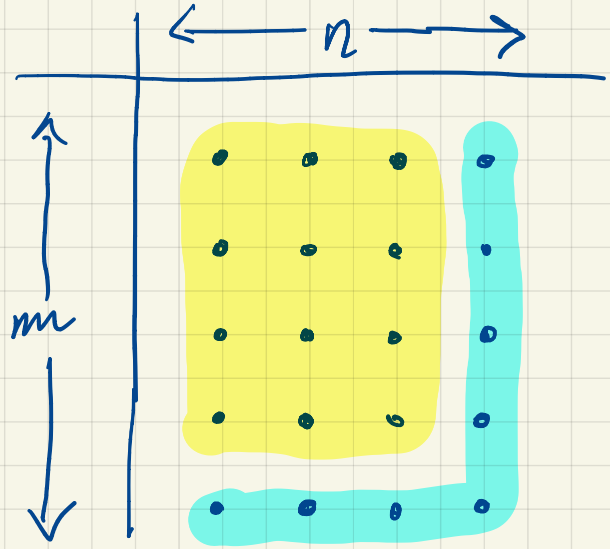
$$\left(\frac{1}{6}, \frac{1}{6}, \frac{1}{3}, \frac{1}{3}\right)$$

$E[AA] = 10 \quad E[AB] = 10 \quad E[BA] = 20 \quad E[BB] = 20$

$$\frac{(14-10)^2}{10} + \frac{(6-10)^2}{10} + \frac{(16-20)^2}{20} + \frac{(24-20)^2}{20} = 4.8$$

Check Pvalue of $\chi^2_{\boxed{?}}$ What should the degree be?

Given $AA = 14$, we can determine all the remaining numbers. Degree of freedom is 1.



In general, knowing yellow can determine blue. So degree of freedom is

$$(m-1)(n-1)$$

Next: χ^2_k as prior ...