G-Test

Important technique in statistics

Allows us to compare an observed categorical distribution to a theoretical/expected one.

	Woods	Cultivated Grassplot	Deciduous Forest	other	Total
Percent Area	4.8%	14.7%	39.6%	40.9%	100%
Counts of Foraging sites	4	16	61	345	426

Does muntjac care where it forages?

If it forages "at random" around Hainan Island, we'd expect that the distribution of the 426 foraging sites would match the distribution of area types.

Expected Counts	$4.8\% \cdot 426 = 20.448$	$14.7\% \cdot 426 = 62.622$	$39.6\% \cdot 426 = 168.696$	$40.9\% \cdot 426 = 174.234$	426

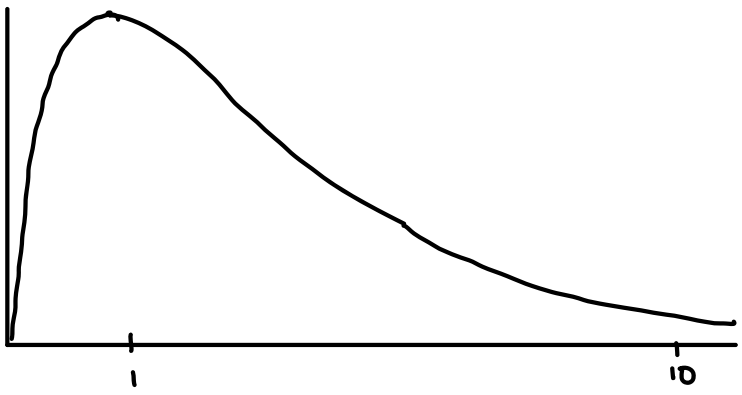
We'll compute a number G that measures "distance" of observed counts from the expected counts.

$$G = 2 \sum O \cdot \ln(O/E)$$

- $G = 0$ when $O = E$, ie, all observed & expected counts are all the same, and

it gets bigger as counts get far from each other.

- (continuing to assume that the expected distribution is "correct", and if expected counts are all "large") the values of G as I sample repeatedly follow a chi-square distribution with n-1 degrees of freedom, where n = # categories. Here n=4, so 3 degrees of freedom.

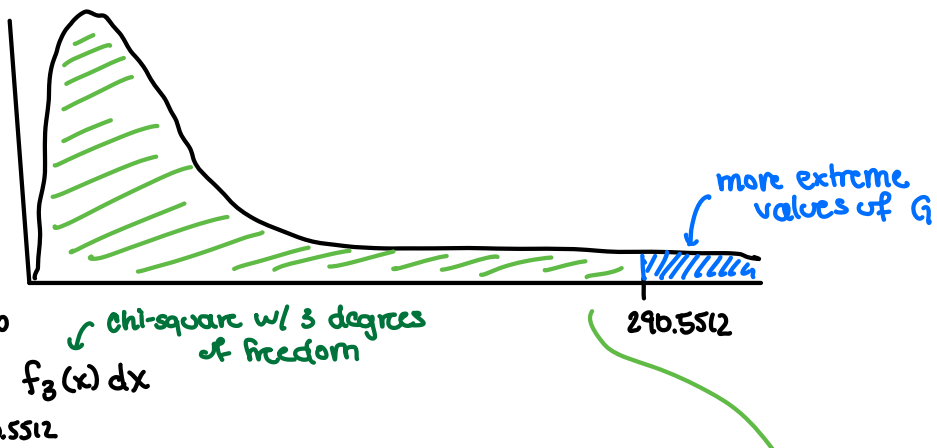


Let's actually compute G

$$G = 2\left(4\cdot\ln\left(\frac{4}{20.448}\right) + 16\cdot\ln\left(\frac{16}{62.622}\right) + 61\cdot\ln\left(\frac{61}{168.696}\right) + 345\cdot\ln\left(\frac{345}{274.284}\right)\right)$$

$$\approx 290.5512$$

This is a very extreme value of G...!



If I compute $\int_{290.5512}^{\infty} f_3(x)\,dx$ that gives me some measure of how unlikely G is.

green area!

$$\text{cum\_distribution\_function}(290.5512) = \int_0^{290.5512} f_3(x)\,dx.$$

The blue area is $1 - \text{cum\_distribution\_function}(290.5512) = 1.1\times10^{-62}$ (tiny!)

If hypothesis that muntjac forages "at random" is correct, a sample like the

one that we saw would be very unlikely!

This suggests strongly that the hypothesis is not correct — muntjac seems to prefer some types of environments on Hainan Island more than others.