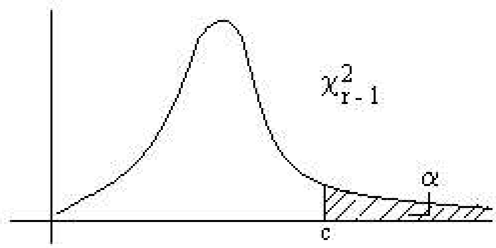
18.05 Lecture 33
May 4, 2005

Simple goodness-of-fit test:
$H_1 : p_i = p_i^0, i \leq r; H_2$: otherwise.

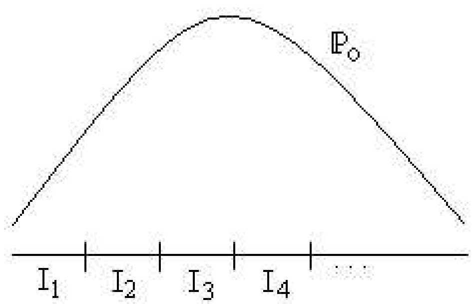
$$\mathcal{T} = \sum_{i=1}^{r} \frac{(N_i - np_i^0)^2}{np_i^0} \sim \chi^2_{r-1}$$



Decision Rule:

$$\delta = \{H_1 : T \leq c; H_2 : T > c\}$$

If the distribution is continuous or has infinitely many discrete points:
Hypotheses: $H_1 : \mathbb{P} = \mathbb{P}_0; H_2 : \mathbb{P} \neq \mathbb{P}_0$

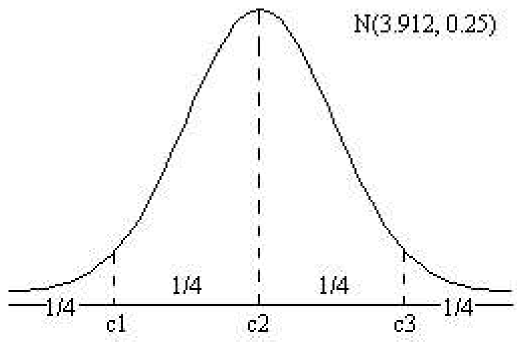


Discretize the distribution into intervals, and count the points in each interval.
You know the probability of each interval by area, then, consider a finite number of intervals.
This discretizes the problem.

New Hypotheses: $H_1' : p_i = \mathbb{P}(X \in I_i) = \mathbb{P}_0(X \in I_i); H_2$ otherwise.
If H_1 is true $\rightarrow H_1'$ is also true.

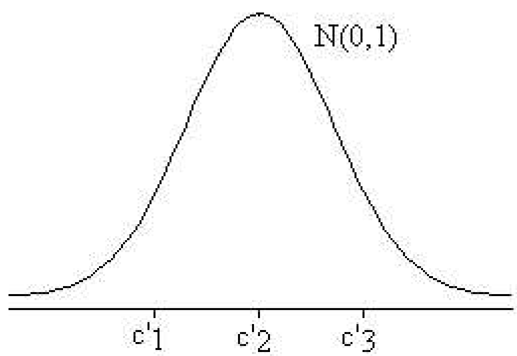
Rule of Thumb:
$np_i^0 = n\mathbb{P}_0(X \in I_i) \geq 5$
If too small, too unlikely to find points in the interval,
does not approximate the chi-square distribution well.

Example 9.1.2 $\rightarrow$ Data $\sim N(3.912, 0.25), n = 23$
$H_1 : \mathbb{P} \sim N(3.912, 0.25)$
Choose k intervals $\rightarrow p_i^0 = \frac{1}{k}$
$n(\frac{1}{k}) \geq 5 \rightarrow \frac{23}{k} \geq 5, k = 4$



$N(3.912, 0.25) \sim X \rightarrow \frac{X-3.912}{\sqrt{0.25}} \sim N(0,1)$
Dividing points: $c_1, c_2 = 3.912, c_3$
Find the normalized dividing points by the following relation:

$$\frac{c_i - 3.912}{0.5} = c_i'$$

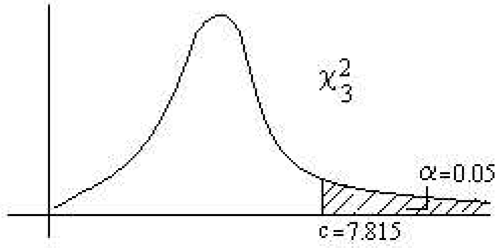


The c_i' values are from the std. normal distribution.
$\rightarrow c_1' = -0.68 \rightarrow c_1 = -0.68(0.5) + 3.912 = 3.575$
$\rightarrow c_2' = 0 \rightarrow c_2 = 0(0.5) + 3.912 = 3.912$
$\rightarrow c_3' = 0.68 \rightarrow c_3 = 0.68(0.5) + 3.912 = 4.249$

Then, count the number of data points in each interval.
Data: $N_1 = 3, N_2 = 4, N_3 = 8, N_4 = 8; n = 23$
Calculate the T statistic:

$$T = \frac{(3 - 23(0.25))^2}{23(0.25} + ... + \frac{(8 - 23(0.5))^2}{23(0.25)} = 3.609$$

Now, decide if T is too large.
$\alpha = 0.05$ - significance level.
$\chi^2_{r-1} \rightarrow \chi^2_3, c = 7.815$



Decision Rule:
$\delta = \{H_1 : T \leq 7.815; H_2 : T > 7.815\}$
$T = 3.609 < 7.815$, conclusion: accept H_1
The distribution is relatively uniform among the intervals.

Composite Hypotheses:
$H_1 : p_i = p_i(\theta), i \leq r$ for $\theta \in \Theta$ - parameter set.
H_2 : not true for any choice of θ

Step 1: Find θ that best describes the data.
Find the MLE of θ
Likelihood Function: $\psi(\theta) = p_1(\theta)^{N_1} p_2(\theta)^{N-2} \times ... \times p_r(\theta)^{N_r}$
Take the log of $\psi(\theta) \rightarrow$ maximize $\rightarrow \widehat{\theta}$

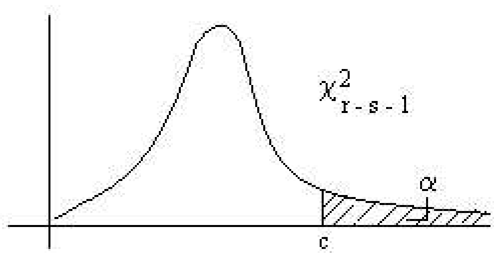
Step 2: See if the best choice of $\widehat{\theta}$ is good enough.
$H_1 : p_i = p_i(\widehat{\theta})$ for $i \leq r, H_2$: otherwise.

$$T = \sum_{i=1}^{r} \frac{(N_i - np_i(\widehat{\theta}))^2}{np_i(\widehat{\theta})} \sim \chi^2_{r-s-1}$$

where s - dimension of the parameter set, number of free parameters.

Example: $N(\mu, \sigma^2) \rightarrow s = 2$
If there are a lot of free parameters, it makes the distribution set more flexible.
Need to subtract out this flexibility by lowering the degrees of freedom.

Decision Rule:
$\delta = \{H_1 : T \leq c; H_2 : T > c\}$
Choose c from χ^2_{r-s-1} with area $= \alpha$



Example: (pg. 543)
Gene has 2 possible alleles A_1, A_2
Genotypes: A_1A_1, A_1A_2, A_2A_2
Test that $\mathbb{P}(A_1) = \theta, \mathbb{P}(A_2) = 1 - \theta$,

but you only observe genotype.

$H_1 : \mathbb{P}(A_1A_2) = 2\theta(1-\theta) \leftarrow N_2$
$\mathbb{P}(A_1A_1) = \theta^2 \leftarrow N_1$
$\mathbb{P}(A_2A_2) = (1-\theta)^2 - \leftarrow N_3$
r = 3 categories.
s = 1 (only 1 parameter, θ)

$$\psi(\theta) = (\theta^2)^{N_1}(2\theta(1-\theta))^{N_2}((1-\theta)^2)^{N_3} = 2^{N_2}\theta^{2N_1+N_2}(1-\theta)^{2N_3+N_2}$$

$$\log\psi(\theta) = N_2\log 2 + (2N_1+N_2)\log\theta + (2N_3+N_2)\log(1-\theta)$$

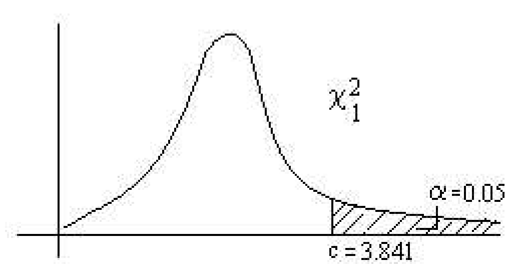
$$\frac{\partial}{\partial\theta} = \frac{2N_1+N_2}{\theta} - \frac{2N_3+N_2}{1-\theta} = 0$$

$$(2N_1+N_2)(1-\theta) - (2N_3+N_2)\theta = 0$$

$$\widehat{\theta} = \frac{2N_1+N_2}{2N_1+2N_2+2N_3} = \frac{2N_1+N_2}{2n}$$

compute $\widehat{\theta}$ based on data.
$p_i^0 = \widehat{\theta}^2, p_2^0 = 2\widehat{\theta}(1-\widehat{\theta}), p_3^0 = (1-\widehat{\theta})^2$

$$\mathcal{T} = \sum \frac{(N_i - np_i^0)^2}{np_i^0} \sim \chi^2_{r-s-1} = \chi^2_1$$



For an $\alpha = 0.05$, c = 3.841 from the χ^2_1 distribution.
Decision Rule:
$\delta = \{H_1 : T \leq 3.841; H_2 : T > 3.841\}$

** End of Lecture 33