CENG 222
Statistical Methods for Computer Engineering

Week 11

Chapter 10

10.1 Chi-square Tests

Chi-square distribution

- Introduced in Section 9.5.1 (not covered)
- Used to model sample variance.
- Recall that sample variance is:
  - $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$
- s^2 is not Normal because the summands $(X_i - \bar{X})^2$ are not independent, they all depend on $\bar{X}$.
- s^2 is also not symmetric (left tail of its distribution ends at 0 because it is always non-negative)

Chi-square distribution

- When X_is are independent and Normal with $\text{Var}(X_i) = \sigma^2$, the distribution of

$$\frac{(n-1)s^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{\sigma}\right)^2$$

is Chi-square with $(n-1)$ degrees of freedom.

- Chi-square (X^2) with *v* degrees of freedom is a continuous distribution with density:

$$f(x) = \frac{1}{2^{\frac{v}{2}}\Gamma(\frac{v}{2})} x^{\frac{v}{2}-1} e^{-x/2}, \qquad x > 0$$

Chi-square distribution

- Chi-square is a special case of Gamma
  - Chi-square(v) = Gamma(v/2,1/2)
  - For example, Chi-square with 2 degrees of freedom is Exponential(1/2)

- Chi-square (X^2) expectation and variance:

$$\mathbf{E}(X) = v$$

$$\mathrm{Var}(X) = 2v$$

- Chi-square (X^2) is introduced by Karl Pearson (1857-1936) who was the teacher of the Student (William Gosset).

Chi-square distribution

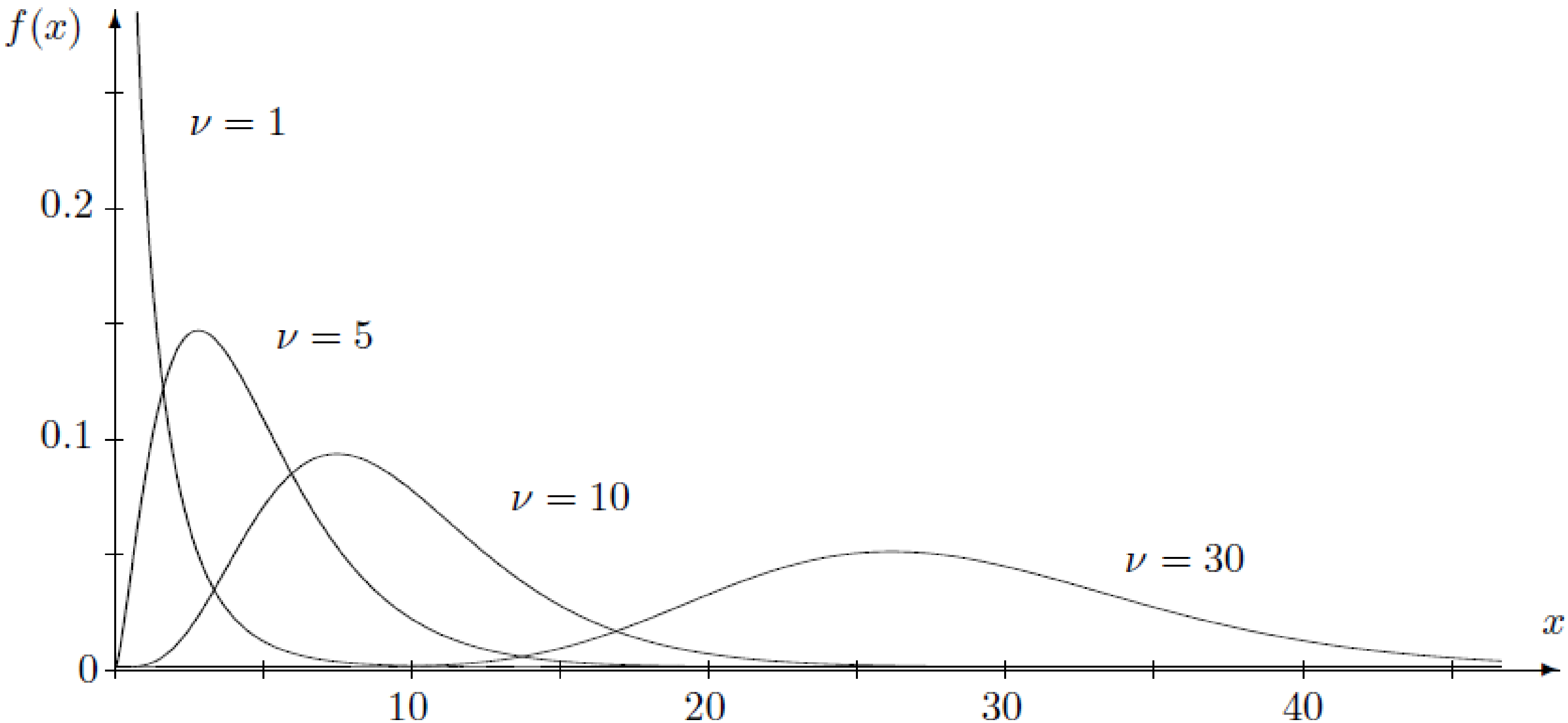

FIGURE 9.12: *Chi-square densities with $\nu = 1, 5, 10,$ and 30 degrees of freedom. Each distribution is right-skewed. For large ν, it is approximately Normal.*

Chi-Square Tests

- Tests of *counts* by comparison of *observed* counts with *expected* counts
  - Use bins for continuous distributions
- Chi-square statistic

$$X^2 = \sum_{k=1}^{N} \frac{\{Obs(k) - Exp(k)\}^2}{Exp(k)}$$

N: number of categories or bins

$Obs(k)$ is the observed counts of sampling units in category k.

$Exp(k) =$ expected number of sampling units is the null hypothesis H_0 is true.

Chi-square tests

- The Chi-square test is always a one-sided right-tail test.
- Level alpha rejection region is:
  - $R = [\mathrm{X}_{\alpha}^{2}, +\infty)$
- P-value is
  - $P = \mathbf{P}(\mathrm{X}^{2} > \mathrm{X}_{obs}^{2})$
- In order to apply Chi-square test, each category should have an expected count of at least 5. If not, merge categories to increase count.

Testing a distribution

- To test whether a sample $(X_1, X_2, \ldots, X_n)$ of size n is from a distribution F_0.
  - H_0 : $F=F_0$ vs H_A : $F \neq F_0$

1. Divide the support of F_0 into bins $B_1 \ldots B_N$ (5-8 bins are sufficient).
2. Count number of sampling units falling into each bin B_k
3. $Exp(k)=nF_0(B_k)$. Check if all expected counts are > 5. If so, compute test statistic and conduct the test; if not, merge bins and restart from Step 1.

Example 10.1: Fair Die?

- 90 tosses of a die are observed

1	2	3	4	5	6
20	15	12	17	9	17

- F_0 = discrete uniform distribution 1..6
- Bins are already defined for this discrete distribution
  - – $Exp(k)$ = 90*1/6=15 (no need to merge bins)
- Compute X^2_{obs}

Example 10.1: Fair Die?

- 90 tosses of a die are observed

1	2	3	4	5	6
20	15	12	17	9	17

- Compute X^2_{obs}

$$X^2_{obs} = \frac{(20-15)^2}{15} + \frac{(15-15)^2}{15} + \frac{(12-15)^2}{15} + \frac{(17-15)^2}{15} + \frac{(9-15)^2}{15} + \frac{(17-15)^2}{15} = 5.2$$

- $v = N - 1 = 5$
- From Table A6, $P = \mathbf{P}(X^2 > 5.2) = 0.2 .. 0.8$
- Cannot reject H_0 . Evidence for unfairness is not sufficient.

Testing a family of distributions

- First, estimate the distribution parameters (may use MLE)
  - Degrees of freedom of X^2 is reduced by the number of distribution parameters estimated
    - $(N - d - 1)$ where d is the number of estimated parameters.
- Then, conduct the X^2 test as before.

Example 10.2: Transmission errors

- Transmission errors in communication channels are usually Poisson. Let's test this.
- 170 channels are randomly selected

0	1	2	3	4	5	7
44	52	36	20	12	5	1

- Estimate lambda
- $\hat{\lambda} = \bar{X} = \frac{44(0)+52(1)+36(2)+20(3)+12(4)+5(5)+1(7)}{170} = 1.55$

Example 10.2: Transmission errors

- 170 channels are randomly selected

0	1	2	3	4	5	7
44	52	36	20	12	5	1

- $\hat{\lambda} = 1.55$
- If we select 6 bins (last bin: # errors $\geq$5) the last bins expected count becomes 3.6. So, reduce to 5 bins (last bin: # errors $\geq$4)

k	0	1	2	3	4
$Exp(k)$	36	55.9	43.4	22.5	12.3
$Obs(k)$	44	52	36	20	18

Example 10.2: Transmission errors

k	0	1	2	3	4
$Exp(k)$	36	55.9	43.4	22.5	12.3
$Obs(k)$	44	52	36	20	18

- $X_{obs}^2 = 6.2$
- $v = N - 1 - 1 = 3$
- From Table A6, $P = \mathbf{P}(X^2 > 6.2) = 0.1 \,..\, 0.2$
- Conclusion: There is no evidence against a Poisson distribution of the number of transmission errors.

Testing independence

- Testing independence of two factors A and B.
- A and B partition the population into k and m categories, respectively.

	B_1	B_2	$\cdots$	B_m	row total
A_1	n_{11}	n_{12}	$\cdots$	n_{1m}	$n_{1\cdot}$
A_2	n_{21}	n_{22}	$\cdots$	n_{2m}	$n_{2\cdot}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
A_k	n_{k1}	n_{k2}	$\cdots$	n_{km}	$n_{k\cdot}$
column total	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot m}$	$n_{\cdot\cdot} = n$

- Use ratios to estimate probabilities $x \in A_i$, $x \in B_j$, and $x \in A_i \cap B_j$

Testing independence

- If the null hypothesis was true, the expected count n_{ij} would be $n\frac{n_{i.}}{n}\frac{n_{.j}}{n}$

	B_1	B_2	$\cdots$	B_m	row total
A_1	n_{11}	n_{12}	$\cdots$	n_{1m}	$n_{1\cdot}$
A_2	n_{21}	n_{22}	$\cdots$	n_{2m}	$n_{2\cdot}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
A_k	n_{k1}	n_{k2}	$\cdots$	n_{km}	$n_{k\cdot}$
column total	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot m}$	$n_{\cdot\cdot} = n$

- $X_{obs}^2 = \sum_{i=1}^{k}\sum_{j=1}^{m}\frac{\{Obs(i,j)-\widehat{Exp}\,(i,j)\}^2}{\widehat{Exp}\,(i,j)}$
- $v = (k-1)(m-1)$

Example 10.4: Spam vs Image Attachments

- A sample of 1000 emails is observed:

$Obs(i,j) = n_{ij}$	With images	No images	$n_{i\cdot}$
Spam	160	240	400
No spam	140	460	600
$n_{\cdot j}$	300	700	1000

- Expected counts are estimated as:

$\widehat{Exp}(i,j) = \frac{(n_{i\cdot})(n_{\cdot j})}{n}$	With images	No images	$n_{i\cdot}$
Spam	120	280	400
No spam	180	420	600
$n_{\cdot j}$	300	700	1000

Example 10.4: Spam vs Image Attachments

- $\mathrm{X}_{obs}^2 = \frac{(160-120)^2}{120} + \frac{(240-280)^2}{280} + \frac{(140-180)^2}{180} + \frac{(460-420)^2}{420} = 31.75$
- $v = (2-1)(2-1) = 1$
- From Table A6, $P < 0.001$.
- We have significant evidence that image attachments are related to being spam.