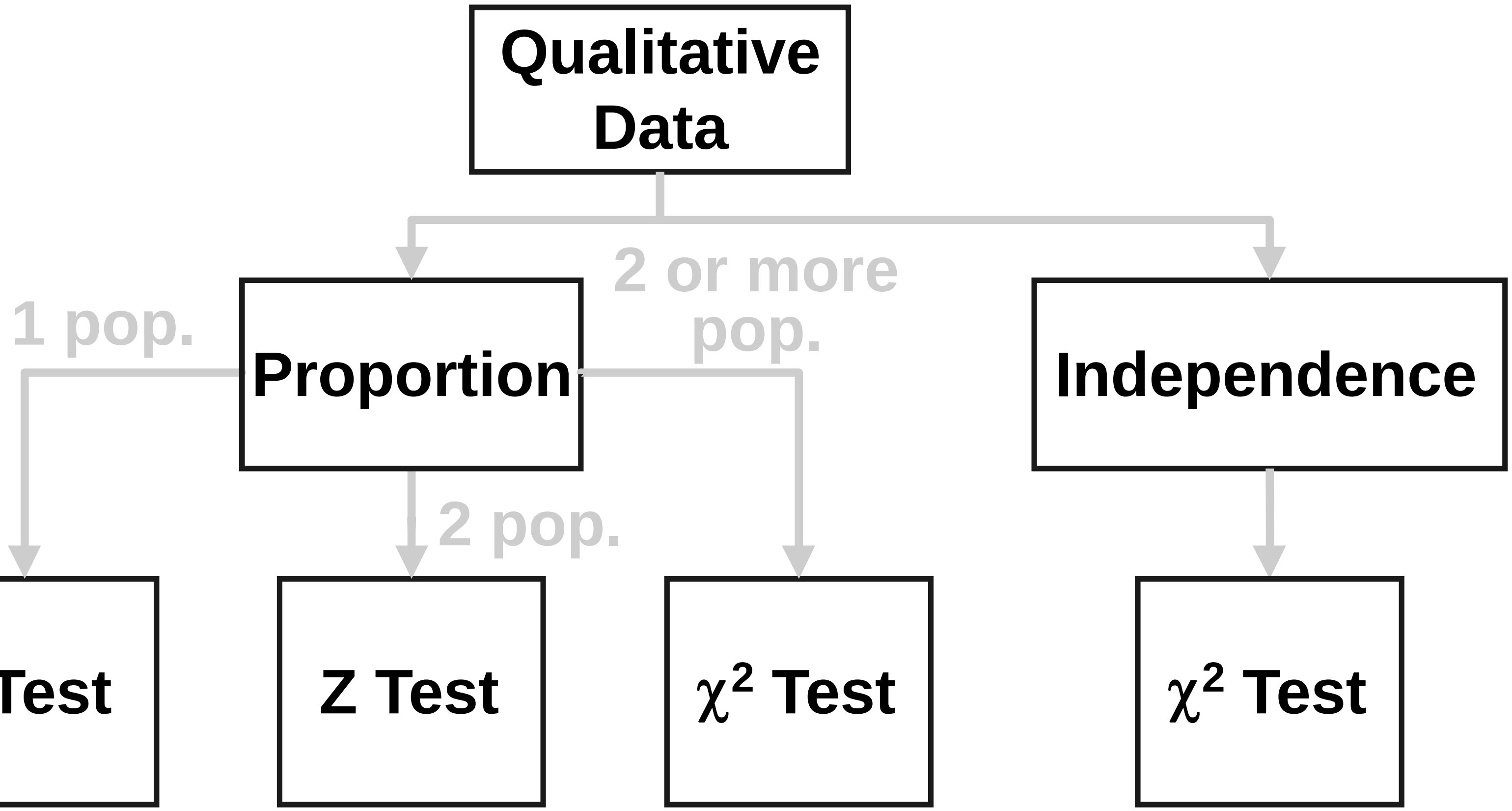
Section 4.7

Chi-Square Tests

Hypothesis Tests for Qualitative Data



Chi-Square Distribution

Theorem : If independent random variables $Z_1,\dots,Z_r \sim N(0,1)$, then

$$Z_1^2+\dots+Z_r^2 \sim \chi^2(r),$$

which is a Chi-Square distribution with degrees of freedom r.

For $X \sim \chi^2(r)$, mean $E(X)=r$, variance $Var(X)=2r$.

Example : $Y_1 \sim Binomial(n, p_1)$. Let $Y_2 = n - Y_1 \sim Binomial(n, 1-p_1)$

(1). $Y_1 - np_1 = (n - Y_2) - np_1 = -(Y_2 - np_2)$, where $p_2 = 1 - p_1$.

$$\frac{1}{np_1}+\frac{1}{np_2}=\frac{p_2+p_1}{np_1p_2}=\frac{1}{np_1p_2}$$

(2). For large n (np ≥ 5, n(1 - p) ≥ 5), normal approximation :

$$\chi^2(1)=Z_1^2 \cong \frac{(Y_1-np_1)^2}{np_1(1-p_1)}=\frac{(Y_1-np_1)^2}{np_1}+\frac{(Y_2-np_2)^2}{np_2}.$$

Chi-Square (χ^2) Test for *k* Proportions

- 1. Tests Equality (=) of Proportions Only
- 2. One Variable With Several Levels
- 3. Assumptions

  (a) Multinomial Experiment (b) All Expected Counts ≥ 5
- 4. Uses One-Way Contingency Table

Multinomial Experiment

- 1. *n* Identical Independent Trials
- 2. *k* Outcomes to Each Trial
- 3. Constant Outcome Probability p_i, i=1,…k, and $\Sigma_i\ p_i = 1$
- 4. Random Variable is Count y_i, i=1,…,k
- 5. Example: Ask 100 people which of 3 candidates they will vote for

Candidate			
Tom	**Bill**	**Mary**	**Total**
35	**20**	**45**	**100**

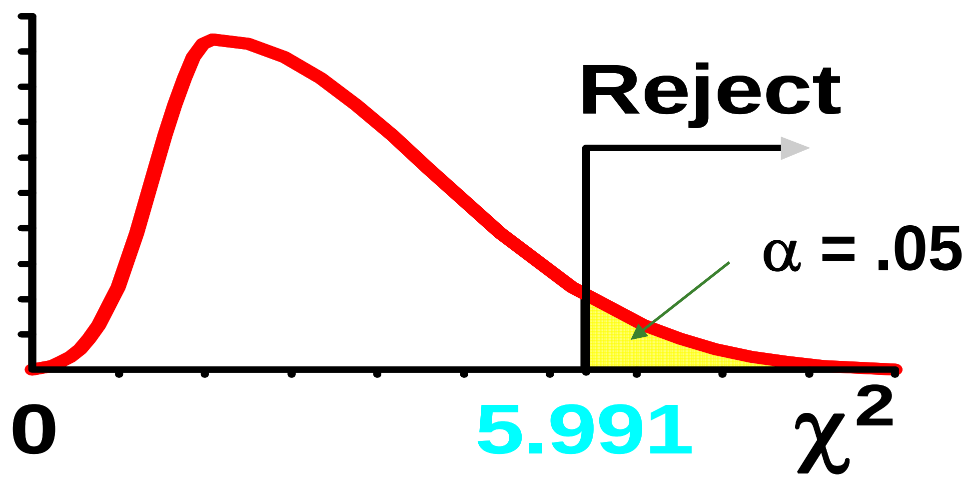
χ^2 Test for k Proportions

- 1. Hypotheses
  - H_0: $p_1 = p_{1,0}$, $p_2 = p_{2,0}$, ..., $p_k = p_{k,0}$
  - H_a: Not all p_i are equal to $p_{i,0}$
- 2. Test Statistic $$\chi^2 = \sum_{i=1}^{k} \frac{\left(y_i - np_{i,0}\right)^2}{np_{i,0}}$$
- 3. Degrees of Freedom under H_0 : df $= k - 1$
- 4. Rejection region $$\left\{\chi^2 > \chi^2_{\alpha}(k-1)\right\}$$

Example

- As personnel director, Mr. A wants to test the perception of fairness of three methods of performance evaluation.
- Of **180** employees, **63** rated **Method 1** as fair. **45** rated **Method 2** as fair. **72** rated **Method 3** as fair.
- At the **.05** level, is there a **difference** in perceptions?

- H0: $p_1 = p_2 = p_3 = 1/3$ vs. Ha: they are different
- $\kappa=3$, $\alpha = .05$, $y_1 = 63$ $y_2 = 45$ $y_3 = 72$
- DF=2, Critical Value: $\chi^2 = 5.991$
- $np_{i,0}=60$, $i=1,2,3$
- Observed test statistic: $\chi^2 = 6.3$



χ^2 Test of Independence

- 1. Shows if a relationship exists between 2 qualitative variables
  - One sample is drawn
  - Does **not** show causality
- 2. Assumptions

  (a) multinomial experiment (b) all expected counts ≥ 5
- 3. Uses two-way contingency table

Observations From 1 Sample Jointly in 2 Qualitative Variables

Levels of variable 2

	House Location		
House Style	**Urban**	**Rural**	**Total**
Split-Level	**63**	**49**	**112**
Ranch	**15**	**33**	**48**
Total	**78**	**82**	**160**

Levels of variable 1

χ^2 Test of Independence (Cont.)

- 1. Testing hypotheses
  - H_0: Variables are independent
  - H_a: Variables are related (or dependent)
- 2. Test Statistic

$$\chi^2 = \sum_{i=1}^{a} \sum_{j=1}^{b} \frac{\left(y_{ij} - n\hat{p}_{i\cdot}\hat{p}_{\cdot j}\right)^2}{n\hat{p}_{i\cdot}\hat{p}_{\cdot j}}$$

Where y_{ij} is the number of observations in cell (i,j) and

$$\hat{p}_{i\cdot} = \frac{y_{i\cdot}}{n} = \frac{1}{n}\sum_{j=1}^{b} y_{ij},\ \hat{p}_{\cdot j} = \frac{y_{\cdot j}}{n} = \frac{1}{n}\sum_{i=1}^{a} y_{ij} \Rightarrow n\hat{p}_{i\cdot}\hat{p}_{\cdot j} = \left(y_{i\cdot} \cdot y_{\cdot j}\right)/n$$

where the row/column total are $y_{i\cdot} = \sum_{j=1}^{b} y_{ij},\ y_{\cdot j} = \sum_{i=1}^{a} y_{ij}$.

- Under null hypothesis (independence), $\chi^2 \sim \chi^2((a-1)(b-1))$.
- Rejection region is $\left\{\chi^2 > \chi^2_\alpha((a-1)(b-1))\right\}$

Example: Chi-Square Test for Independence

- In one large factory, 100 employees were judged to be highly successful and another 100 marginally successful.
- All workers were asked, "Which do you find more important to you personally, the money you are able to take home or the satisfaction you feel from doing the job?"
- In the first group, 49% found the money more important, but in the second group 53% responded that way.
- Test the null hypothesis that job performance and job motivation are *independent* using the .01 level of significance.

	High Success	Marginal Success	Total
Money	**49**	**53**	**102**
Satisfaction	**51**	**47**	**98**
Total	**100**	**100**	**200**

Goodness-of-fit Test

- A population X may follow a distribution with one or two parameters
- Divide outcome space into k mutually exclusive and exhaustive cells, then decide the frequencies of those cells, y_i, i=1,…,k, and $\Sigma y_i = n$
- Expected frequency (probability) of each cell p_i are determined by the population distribution if the parameters are specified.
- Assumption: expected counts of each cell $np_i \geq 5$
- Hypotheses
  - H_0: X follows a distribution (Normal, Poisson, etc.)
  - H_a: X does not follow the specified distribution

Then

$$\chi^2 = \sum_{i=1}^{k} \frac{(y_i - np_i)^2}{np_i} \sim \chi^2(k-1-h) \text{ under } H_{0.}$$

where degrees of freedom is **(*k*-1-*h*)** and ***h*** is the number of unknown parameters specified in null hypothesis.

Rejection Region: $\left\{\chi^2 > \chi^2_\alpha(k-1-h)\right\}$

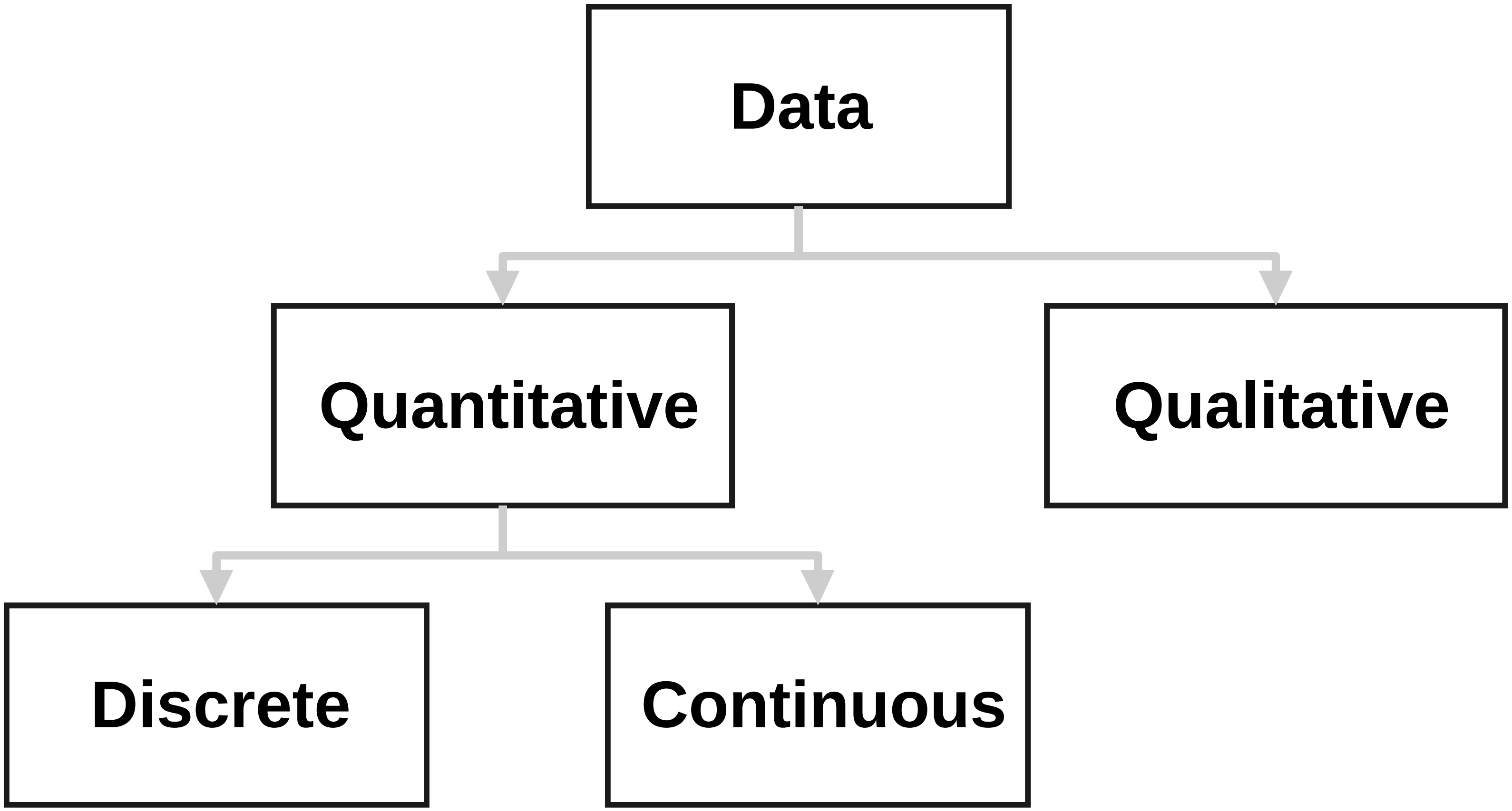
Example: We observe n=85 values of a r.v. X that is thought to have a Poisson distribution

x	0	1	2	3	4	5
Frequency	41	29	9	4	1	1

Hypotheses:

$$H_0 : X \sim Poisson(\lambda) \quad \text{vs} \quad H_1 : X \text{ does not follow } Poisson \text{ distribution}$$

Data Types



R code: F-test and t-test

```
>sole <- read.table("H:/Teaching/STAT-481/sole.txt", header=TRUE, sep="\t")
>names(sole)
[1] "Boy" "MA"  "MB"

> var.test(sole$MA, sole$MB)              ## test of equal variance
>   F test to compare two variances
    data:  sole$MA and sole$MB
F = 0.9474, num df = 9, denom df = 9, p-value = 0.9372
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
       0.2353191   3.8142000
sample estimates:
ratio of variances
       0.9473933

>  ##  Use two-sample t-test with equal variances
```

```
>t.test(sole$MA, sole$MB, var.equal=T)
>       Two Sample t-test
         data:  sole$MA and sole$MB
t = -0.3689, df = 18, p-value = 0.7165
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
          -2.744924  1.924924
sample estimates:
        mean of x mean of y
         10.63     11.04

> ## paired comparison design  --  two-tailed test

> t.test(sole$MA, sole$MB, paired=T)
 >t.test(sole$MA, sole$MB, paired=T)$statistic
> t.test(sole$MA, sole$MB, paired=T)$p.value
> t.test(sole$MA, sole$MB, paired=T)$conf.int

 # right-tailed paired t-test
  >t.test(sole$MA, sole$MB, paired=T, "greater")
```

Normality Check (R code)

```
x.norm <- rnorm(n=100, m=5, sd=1)                 ## Normal distribution mean=5, var=1

boxplot(x.norm, main="Boxplot")                   ## Boxplot
hist(x.norm, main="Histogram of the data")        ## Histogram

plot(density(x.norm), main="Density estimate")    ## Density Estimate
qqnorm(x.norm)                                    ## QQ-plot

z.norm <- (x.norm - mean(x.norm))/sd(x.norm)      ## standardization
qqnorm(z.norm)                                    ## QQ-plot of z.norm
abline(0, 1)                                      ## Add a straight line: y = a + b*x

ks.test(z.norm, "pnorm",  m=0, sd=1)              # One-sample Kolmogorov-Smirnov test

shapiro.test(x.norm)                              #  Shapiro-Wilk normality test
```

Chi-Square test for k Proportions (R code)

```
method =1:3
k = 3
count = c(63, 45, 72)
n = sum(count)
data = cbind(method, count)

## expected probability / expected count
p0 = c(1/3, 1/3, 1/3)
count.exp = n*p0

## observed chisquate test statistic
chisq.obs <- sum((count - count.exp)^2/count.exp)

## p-value
1 - pchisq(chisq.obs, df=k-1)

## rejection region given level=alpha
alpha <- 0.05
chisq.obs > qchisq(1-alpha, df=k-1)
```

Chi-Square test for Independence (R code)

```
raw <- c(49, 53, 51, 47)  ;  n <- sum(raw)
data <- matrix(raw, 2, 2, byrow=TRUE)                ## read data in matrix
a <- ncol(data)  ;  b <- nrow(data)

# cell averages
row.tot <- apply(data, 1, sum)                       ## row sum ##
p.idot <- as.vector(row.tot/n)
col.tot <- apply(data, 2, sum)                       ## column sum ##
p.doti <- as.vector(col.tot/n)

# cell expected averages under independence
cellprob.exp <- (p.idot) %*% t(p.doti)               ## '%*%' matrix product  ##
cellmean.exp <- n*cellprob.exp

# Observed Chisquare Test Statistic
chisq.obs <- sum( (data - cellmean.exp)^2/cellmean.exp)

1 - pchisq(chisq.obs, df = (a-1)*(b-1) )             ##   p-value

alpha <- (0.01)                                      ## significance level
chisq.obs > qchisq( 1 – alpha, df = (a-1)*(b-1) )
```