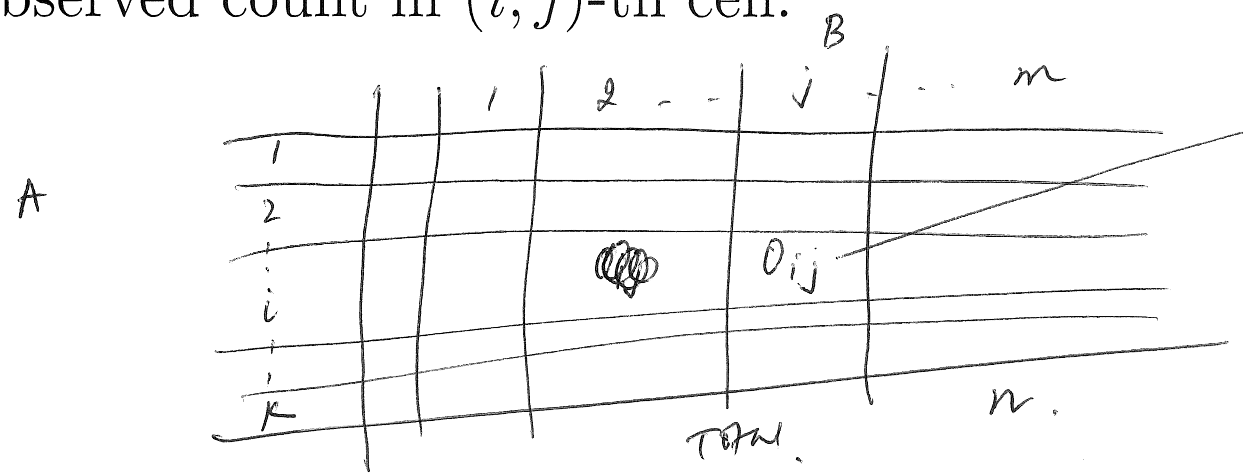
Chi-square test of independence

Set up: Count data on two categorical variables (or factors) A and B obtained from a sample of n subjects. Suppose the categories of A are $i = 1, \ldots, \cancel{m}\,k$, and the categories of B are $j = 1, \ldots, m$. The data are arranged in a $k \times m$ table. Let O_{ij} = observed count in (i, j)-th cell.



$\underbrace{P[X=i, Y=j]}$ $= p_{ij}$

prob. that a randomly selected individual falls in ith category of A and jth category of B

Hypotheses: $H_0 : A$ and B are independent (i.e., are not associated), vs., $H_1 : A$ and B are not independent (i.e., are associated). If there is an association, the value one variable depends (at least to some extent) on the value of the other.

Example: The table below shows 695 children under 15 years of age are cross-classified according to ethnic group and hemoglobin level. Is hemoglobin level associated (related) to ethnicity?

$P[X=1, Y=1] = P[X=1] \cdot P[Y=1]$ indep.

Y

Ethnic Group (X)	Hemoglobin Level (g/100 ml): ≥ 10 (Gr 1)	9.0 - 9.9 (Gr 2)	< 9.0 (Gr 3)	Total	Proportion
A – Gr 1	80	100	20	200	200/695 – $\hat{P}(X=1)$
B – Gr	99	190	96	385	385/695 – $\hat{P}(X=2)$
C – Gr 3	70	30	10	110	110/695 – $\hat{P}(X=3)$
Total	249	320	126	695	
Proportion	249/695	320/695	126/695		

$\hat{P}[X=1]$

- If He level is not associated to ethnicity, then the proportion of subjects in population that fall a He group does not depend on ethnicity, i.e., it is the same for each ethnicity group, and vice versa.

$P[\text{He group}=1 \mid \text{ethnicity}=A] = P[\text{"} \mid \text{ethnicity}=B] = P[\text{"}=\text{"} \mid \text{"}=C]$

This relationship should hold for all He groups. $P[\text{He group}=1]$

H_0: The two variables are indep. , H_1: NOT indep.

To do a chi-square test, we need the expected counts E_{ij} assuming that H_0 is true. Let X and Y indicate respective categories of A and B in which a randomly selected subject from the population falls. When A and B are independent,

$$P(X = i, Y = j) \overset{H_0}{=} P(X = i)P(Y = j) \text{ for all } i, j.$$

- $P(X = i)$ is estimated as $\frac{i\text{-th row total}}{n} = \frac{\sum_l O_{il}}{n}$
- $P(Y = j)$ is estimated as $\frac{j\text{-th column total}}{n} = \frac{\sum_l O_{lj}}{n}$
- Assuming independence, $P(X = i, Y = j)$ is estimated as $\hat{P}(X=i) \cdot \hat{P}(Y=j)$
- Assuming independence, E_{ij} is estimated as

$$\hat{E}_{ij} = (n)\, \hat{P}(X=i) \cdot \hat{P}(Y=j) = \frac{(i\text{-th row total})(j\text{-th col. total})}{n}$$

Test statistic:

$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}} \sim \chi^2_\nu$$ when all $\hat{E}_{ij} \geq 5$. when H_0 is true.

Degrees of freedom:

Read the book and verify! $\nu = (k-1)(m-1)$

categories of A # categories of B.

Example (continued): The expected counts for all cells (in parenthesis below next to the observed counts) are:

	Hemoglobin Level (g/100 ml)			
Ethnic Group	≥ 10	9.0 - 9.9	< 9.0	Total
A	80 (?)	100 (92.09)	20 (36.26)	200
B	99 (137.94)	190 (177.27)	96 (69.80)	385
C	70 (39.41)	30 (50.65)	10 (19.94)	110
Total	249	320	126	695

$$? = \frac{(200)(249)}{695} = ?$$

$$\chi^2 = 67.8$$

verify

R code:

```
> x <- c(80, 100, 20, 99, 190, 96, 70, 30, 10)
> xmat <- matrix(x, byrow=T, ncol=3)
> xmat
     [,1] [,2] [,3]
[1,]   80  100   20
[2,]   99  190   96
[3,]   70   30   10
> chisq.test(xmat)

        Pearson's Chi-squared test

data:  xmat
X-squared = 67.8015, df = 4, p-value = 6.606e-14

>
```

Take $\alpha = 0.05$ or $\alpha = 0.01$

$(k-1)(m-1)$

Conclusion: Reject H_0

$\Rightarrow$ Conclude that the two variables are associated.

Cannot deduce any causal relationship - just that the two variables are related.

Chi-Square test of Homogeneity

Often we are interested in comparing different populations with respect to a variable of interest, e.g., are the populations of carriers and non-carriers of a certain antigen *homogeneous* with respect to blood type?

The dist. of blood group type is same for two populations.

Example: A sample of 150 carriers of a certain antigen and a sample of 500 non-carriers showed the following blood group distributions:

~~with 3 populations~~ Ho: $p_1 = p_2 = p_3$

$P[BG = O \mid \text{carrier}] = P[BG = O \mid \text{Non carrier}] = P[BG = O]$

This equality should also hold for all other BG types.

Blood Group	Carriers	Non-Carriers	Total
O	72	230	302
A	54	192	246
B	16	63	79
AB	8	15	23
Total	150	500	650

NOTE: Test of homogeneity is a generalization of two-sample z test for comparing two proportions

Find p-value.

Are carriers and non-carriers similar with respect to blood group distributions?

→ Do a chisquare test of independence, and we know how to do it.

Test of Homogeneity vs. Test of Independence

Comparing the layout of this table with the table for the test of independence, we see that the two layouts are
Thus, mathematically the tests of homogeneity and independence are exactly the same. So, the same formulas apply. However, there are some key conceptual differences.

Sampling procedure:

- *Test of independence*: <u>one</u> overall sample is collected first and then each observation is classified by levels of the two variables. So, neither row nor column totals are fixed in advance.
- *Test of homogeneity*: <u>several</u> samples are collected from several populations with each sample size fixed in advance. After collecting these pre-determined # of observations, each is classified by various levels of one variable. So, in the above example, column .... totals are fixed.

Number of variables:

- *Test of independence*: **two** variables.
- *Test of homogeneity*: **one** variable. The column/row representing "population" is fixed due to the sampling process.

→ Blood group type.

Hypotheses:

- *Test of independence*: H_0: A and B are indep.
- *Test of homogeneity*: H_0: ~~Tho~~ The populations are the same w.r.t. the variable of interest.

↑ Question of independence doesn't arise because only one variable is being measured.

Recall: X and Y are indep. if

$P[X=i, Y=j] = P[X=i] \cdot P[Y=j]$ for all i,j

This condition is equivalent to:

$P[X=i \mid Y=j] = P[X=i]$ for all i,j.

or

$P[Y=j \mid X=i] = P[Y=j]$ for all i,j.

Nonparametric Tests

Issue: We would like test hypothesis on **center** of a distribution (one-sample problem) or compare centers of two distributions (two-sample problem). But the distributions are not normal — e.g., they are skewed or data has outliers. mean is not a good meas. of center.

Q: Why not simply use large-sample z test?

Fine if mean is a good measure of center

Alternative measure of "center": median of dist.

Nonparametric procedures:

- Typically they don't assume a specific distributional form (e.g., normal); only that the distribution is continuous. X is a cont. r.v. Some procedures assume that the distribution is symmetric.
- More broadly applicable than **parametric procedures** that assume specific distributional form.
- Use these when the distributional assumption behind a parametric procedure is clearly violated.

Sign test

If X is cont.
$P[X = M_0] = 0$.

Data: $X_1, \ldots, X_n$ — i.i.d. sample from X.

pop. median; null value.

Hypotheses: $H_0 : M = M_0$ vs. H_1 : one of three possibilities, $M > M_0$ or $M < M_0$ or $M \neq M_0$

Signs: Remove the X's that are equal to M_0 and reduce the sample size accordingly. — $n^* = \#$ obs. that are not equal to M_0.

sign of $X_i = \begin{cases} +1, & X_i > M_0 \\ -1, & X_i < M_0. \end{cases}$

'-1' $+1$ M_0 sorted x's

Test statistic:

$S = \#$ positive signs.

Null distribution:

If H_0 is true: sign of $X_i \sim \text{Bernoulli}(p=\frac{1}{2})$

$\Rightarrow S \sim \text{Bin}(n^*, p=\frac{1}{2})$.

When to reject H_0?

p-value

- $H_1 : M > M_0$: S is too large (compared with $\frac{n^*}{2}$) $P[\text{Bin}(n^*,\frac{1}{2}) \geq S_{obs}]$
- $H_1 : M < M_0$: S is too small (" " $\frac{n^*}{2}$) $P[\text{Bin}(n^*,\frac{1}{2}) \leq S_{obs}]$
- $H_1 : M \neq M_0$: S is either too large or too small.

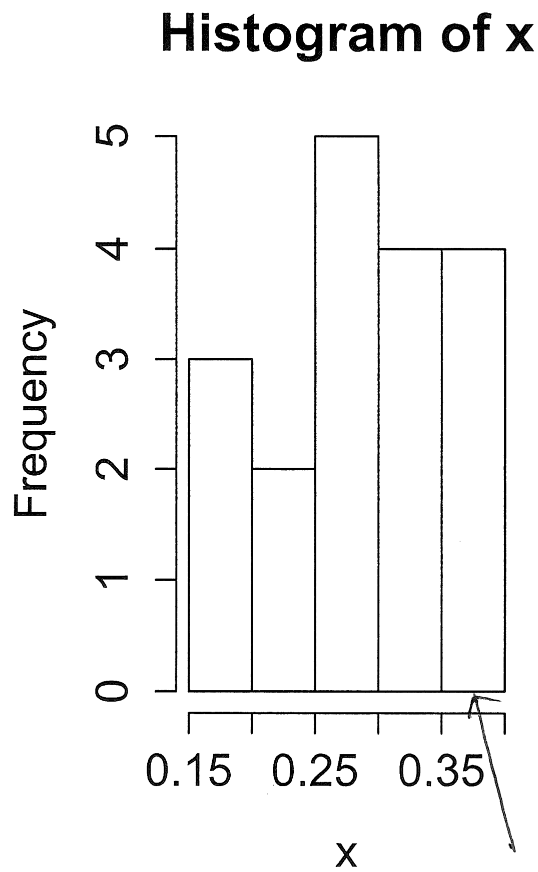
$2 \min\{$the two one-sided p-values$\}$.

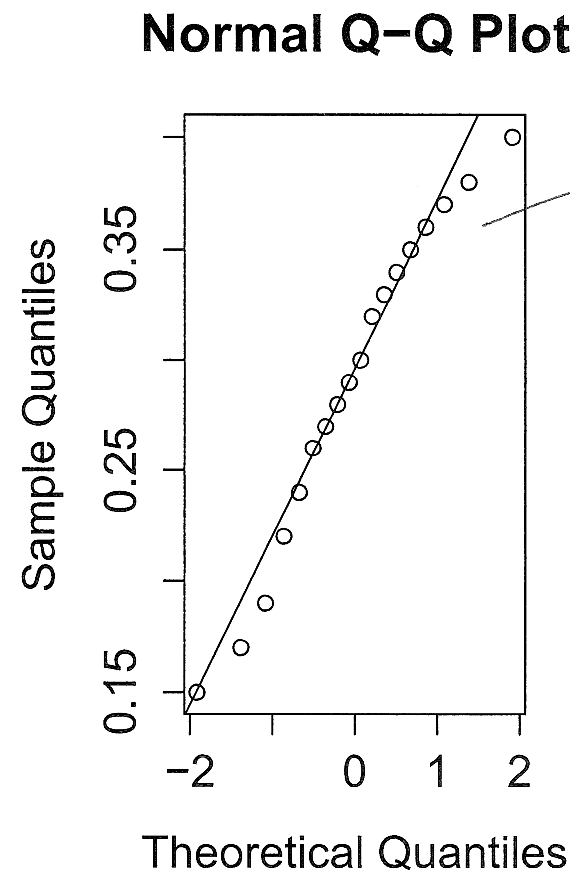
R code:

```r
# Time between keystrokes data from Example 10.9

x <- c(0.24, 0.22, 0.26, 0.34, 0.35, 0.32, 0.33, 0.29,
       0.19, 0.36,
       0.30, 0.15, 0.17, 0.28, 0.38, 0.40, 0.37,0.27)


# Histogram and boxplot

par(mfrow=c(1,2)) # 2 plots in 1 row

hist(x)
qqnorm(x)
qqline(x)

library(nortest)
```

```
> shapiro.test(x)

Shapiro-Wilk normality test

data:  x
W = 0.9611, p-value = 0.6233
>
```

Normality appears reasonable from this as well Q-Q plot.





normality appears reasonable.

suggests uniform distribution.

NB Both models may be reasonable for these data.
With a larger sample size, we may be able to
better distinguish b/w the two distributions.