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?

	Hemoglobin Level (g/100 ml)				
Ethnic Group	≥ 10	9.0 – 9.9	< 9.0	Total	Proportion
A	80	100	20	200	200/695
B	99	190	96	385	385/695
C	70	30	10	110	110/695
Total	249	320	126	695	
Proportion	249/695	320/695	126/695		

- 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.

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?

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

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

same table format as in χ^2 test of independence.

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

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*: one 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*: several 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, ... 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.

Hypotheses:

- *Test of independence*: H_0: Two variables are indep.
- *Test of homogeneity*: H_0: The populations are identical wrt the one variable of interest.

R code:

```
x <- c(72, 230, 54, 192, 16, 63, 8, 15)
xmat <- matrix(x, byrow=T, ncol=2)
# > xmat
#      [,1] [,2]
# [1,]   72  230
# [2,]   54  192
# [3,]   16   63
# [4,]    8    1
# >
# > chisq.test(xmat)
# Pearson's Chi-squared test

# data:  xmat
# X-squared = 2.4052, df = 3, p-value = 0.4927
# >
```

levels of the variable

$df = (4-1) \times (2-1)$

populations

$= (4-1) \times (2-1) = 3.$

$$\sum_{i=1}^{4} \sum_{j=1}^{2} \frac{(O_{ij} - E_{ij})}{E_{ij}}$$

Accept H_0

$\Rightarrow$ Two pop. are homogeneous

$\Rightarrow$ The blood group type dist. among the carriers and non-carriers are same

p-value = $P[\chi^2_3 \geq 2.4052]$ =

R

1 - pchisq(2.4052, 3)

↑ calculate this and verify.

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.

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

Alternative measure of "center":

Nonparametric procedures:

- Typically they don't assume a specific distributional form (e.g., normal); only that the distribution is continuous. 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.

One-sample problem

M = median of X

Sign test

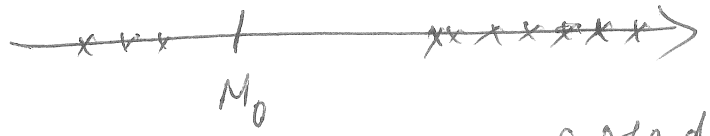
note: no assumptions about the shape of the dist. of X

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

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. (Let $n^* = \#$ obs. that are not equal to M_0.)

Test statistic: $S = \#$ positive signs



sorted obs.

Null distribution: $S \sim \text{Binomial}(n, p = \frac{1}{2})$

When to reject H_0?

- $H_1 : M > M_0$: reject H_0 if S is too large (compared to what we expect, which is $\frac{n}{2}$)
- $H_1 : M < M_0$: reject H_0 if S is too small
- $H_1 : M \neq M_0$: reject H_0 if S is either too large or too small

p-value $= P[\text{Bin}(n, \frac{1}{2}) \geq \text{obs}]$

p-value $= P[\text{Bin}(n, \frac{1}{2}) \leq \text{obs}]$

p-value $= 2 \times \min\{\text{p-values of the two one-sided tests}\}$

Associate a sign to X:

$$X^* = \begin{cases} 1, & X > M_0 \\ 0, & X < M_0 \end{cases}, \qquad P[X^* = M_0] = 0.$$

$$X^* \sim \text{Bernoulli}\left(P = \underbrace{P(X > M_0)}_{\text{"}}\right)$$

$\frac{1}{2}$ if H_0 is true.

Now: Associate a sign to X_i:

$$X_i^* = \begin{cases} 1, & X_i > M_0 \\ 0, & X_i < M_0 \end{cases}$$

$$\Rightarrow X_1^*, \ldots, X_n^* \overset{\text{random sample}}{\sim} \text{Bernoulli}\left(P = \underbrace{P(X > M_0)}_{\text{"}}\right)$$

$\frac{1}{2}$ if H_0 is true.

Define:

$$S = \sum_{i=1}^{n} X_i^* = \#\text{ obs. in the sample that are } > M_0$$

$$= \#\text{ of } \oplus \text{ve signs in sample}$$

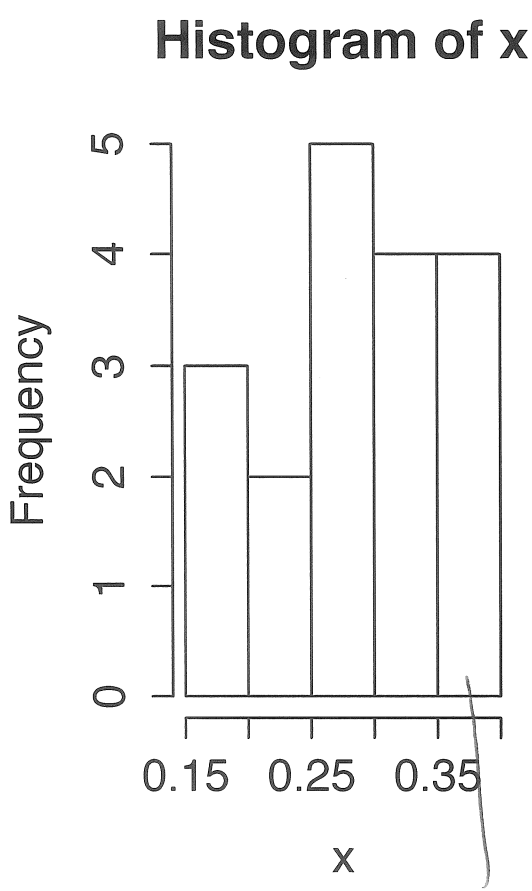
$$\sim \text{Binomial}\left(n, P = \underbrace{P(X > M_0)}_{\text{"}}\right)$$

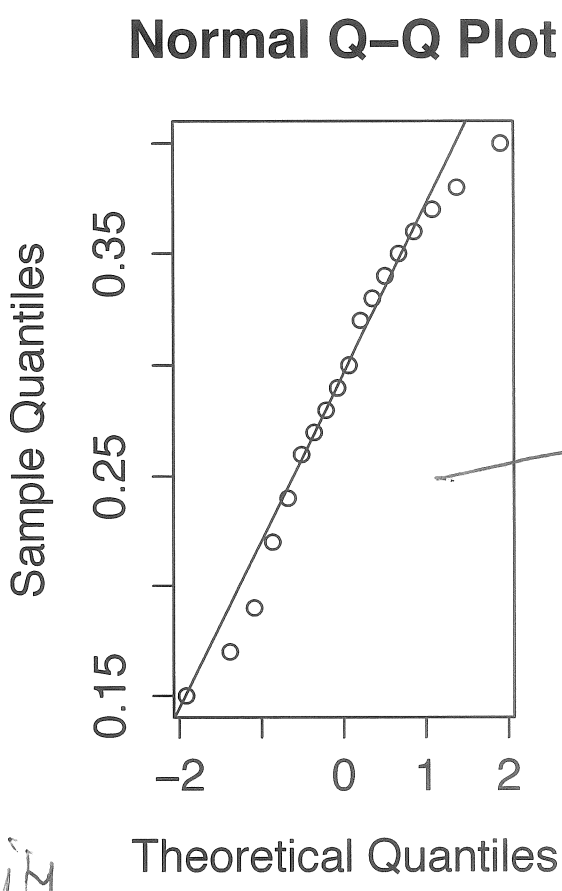
$\frac{1}{2}$ if H_0 is true.

R code:

```
# 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 OK

normality NOT OK

The situation is not so clear-cut because of the size of n.

Let's apply the non-parametric also and see what we conclude.

known value for the n/c owner

```
# Sign test of H0: M = 0.2 vs M is not equal to 0.2
```

$H_1: M = 0.2$

```
sign.stat <- sum(x > 0.2)
```

Obs. > 0.2

```
> sign.stat
[1] 15
>
```

for doing sign test

```
> binom.test(sign.stat, n=sum(x != 0.2), p = 0.5,
alt="two.sided", conf.level=0.95)

Exact binomial test
```

reject H_0 Evidence of unauthorized attempt.

```
data:  sign.stat and sum(x != 0.2)
number of successes = 15, number of trials =
  18, p-value = 0.007538
alternative hypothesis: true probability of
success is not equal to 0.5
```

verify that it matches with obs.

```
95 percent confidence interval:
 0.5858225 0.9642149
sample estimates:
  probability of success
0.8333333
>
```

→ for $p = P[X > \mu]$.