MAT20306 - Advanced Statistics

Lecture 3: Binomial Test (one population proportion)
Fisher Test (two population proportions)

WAGENINGEN UR
For quality of life

Biometris
Quantitative Methods brought to Life

Inference about one or two population proportions

	Situation
1. 1 binary variable : One proportion / probability	10
2. Two proportions / probabilities	11

For both situations we will discuss for the parameter of interest:

- how to construct a 2-sided *approximate* (1-α) confidence interval using a z-approximation.
- How to perform an exact test (Binomial test, Fisher exact test)

Study guide scheme

II Situations where Normality is not assumed (because it does not seem to be appropriate)

	Situation description	Parameter(s) / Questions	Inference	Name / Type of test	Lecture	O&L
		Inference based on ranks of (a) numerical, continuous, variable(s)				
1a	**1 random sample, 1 quantitative response**	Population median. H_0: median = m	T	sign test (or: Wilcoxon signed rank test for $d_i = y_i - m$)	-	5.9
2a	**1 random sample, quantitative responses *x* and *y*, paired data**	Systematic difference between distributions of *x* and of *y*?	T	Wilcoxon signed rank test for $d_i = x_i - y_i$	2	6.5
3a	**2 independent samples/ CRD with** 2 **treatments, 1 quantitative response**	Systematic difference in *y* between the 2 sub-populations/ treatments? Shift alternative.	T	Wilcoxon rank sum test (Mann-Whitney test)	2-3	6.3
4a	**1 quantitative response *y*, 1 qualitative factor** (random samples from t sub-populations or CRD with t treatments (t >2).	Systematic differences in distribution of *y* between the treatments? Shift alternative.	T	Kruskal-Wallis test	8	8.6
		Inference for binary data and categorical data				
10	**1 random sample, binary variable X** *Model:* $P(X_i=1) = \pi$, $P(X_i=0) = 1 - \pi$, $i=1..n$	population fraction or success probability π	E, CI T	z-procedure Binomial test (SPSS / PQRS)	3	10.2
11	**2 independent samples, binary variable X** *or* 2 treatments with CRD	$\pi_1 - \pi_2$: difference in pop. fraction or success probability between sub-populat./treatments	E, CI T	z-procedure. Fisher's exact test (SPSS / PQRS)	3	10.3
12	**1 random sample, 1 nominal variable** (variable with outcomes in *k* classes)	$\pi_1, \pi_2, \ldots, \pi_k$ H_0: $\pi_1 = \pi_{10}, \pi_2 = \pi_{20}, \ldots, \pi_k = \pi_{k0}$	E, T	Pearson's chi-square test for goodness of fit	4	10.4
13	**1 random sample, 2 nominal variables** (outcomes in contingency table with *r* rows and *c* columns)	π_{ij} ($i = 1\ldots r$; $j = 1\ldots c$), probabilities in one population H_0: $\pi_{ij} = \pi_{i.} * \pi_{.j}$	E, T	chi-square test for independence	4	10.5
14	***r* samples, 1 nominal variable with *c* classes** (outcomes in contingency table with *r* rows and *c* columns)	π_{ij} ($i = 1\ldots r$; $j = 1\ldots c$), probabilities per population H_0: $\pi_{11} = \pi_{21} = \ldots = \pi_{r1}$ … $\pi_{1c} = \pi_{2c} = \ldots = \pi_{rc}$	E, T	chi-square test for homogeneity	4	10.5

An example of one proportion – alcohol abuse O&L Example 10.5

The proportion of binge drinking among students in an extensive survey is 0.44. On a large university n = 500 randomly selected students were asked if they engage in binge drinking.

1. Is the proportion (π) of students at that university that engage in binge drinking larger than 0.44?

Population of interest is population of students at the university.
Sampling units are students.

Response x: student is a binge drinker (x = 1) or not (x = 0).
This response variable x is a **binary variable.** → situation 10

Inference about a population proportion

How to estimate π? π is:

1) the population proportion of binge drinkers
2) the probability that a randomly drawn student is a binge drinker
3) the population mean of **x**: drinker(x = 1) or healthy(x = 0).

Define $y = \sum_{i=1}^{n} x_i$ = number of successes in a sample (of size n)

The sample mean is $\frac{y}{n}$, the sample proportion of binge drinkers.

The estimator $\widehat{\pi}$ for π is the sample proportion $\frac{y}{n}$

$$\hat{\pi} = y / n$$

The population proportion (π) of binge drinkers is estimated by the sample proportion ($\widehat{\pi}$) of binge drinkers

Situation 10: 1 probability / fraction

Data: 1 sample $x_1, \ldots, x_n$ with binary outcomes (1 or 0).

Model: the x_i's, i=1,..,n, are independent randomly sampled, all with equal probability of being a 'success' $\pi = P(x=1)$

Parameter of interest: (the population fraction of successes)

$$\pi \quad (= \mu_x)$$

Estimator (method of estimation) : (y = number of successes in the sample.)

$$\hat{\pi} = y / n \quad (= \bar{x})$$

Standard error of the estimator (CI calculation)

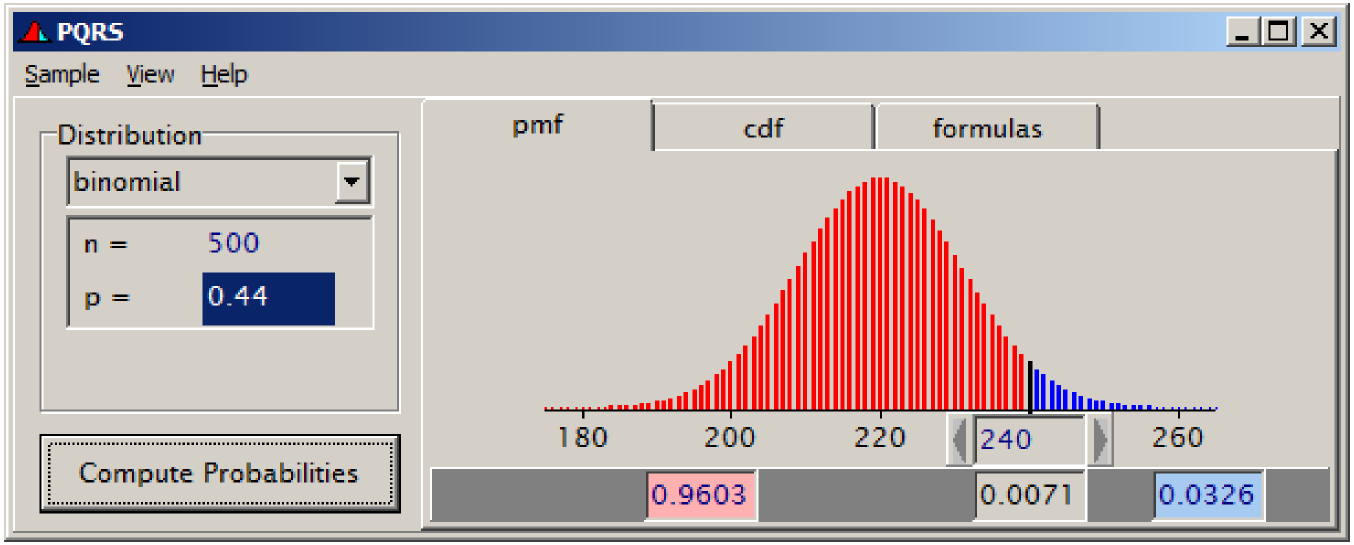
$$se(\hat{\pi}) = \sqrt{\hat{\pi}(1-\hat{\pi})/n}$$

Relevant distribution:

$$y \sim \text{Binomial}(n, \pi)$$

Binomial Distribution

- number of successes y ~ Binomial(n, π)



$E(y) = n\pi$. $\quad$ $\mathrm{Var}(y) = n\pi(1-\pi)$.

variance is maximal when $\pi = ½$

variance vanishes to 0 when $\pi \to 0$ or $\pi \to 1$.

An example of one proportion – alcohol abuse O&L Example 10.5

The proportion of binge drinking among students in an extensive survey is 0.44. On a large university n = 500 randomly selected students were asked if they engage in binge drinking.

1. Is the proportion (π) of students at that university that engage in binge drinking larger than 0.44?

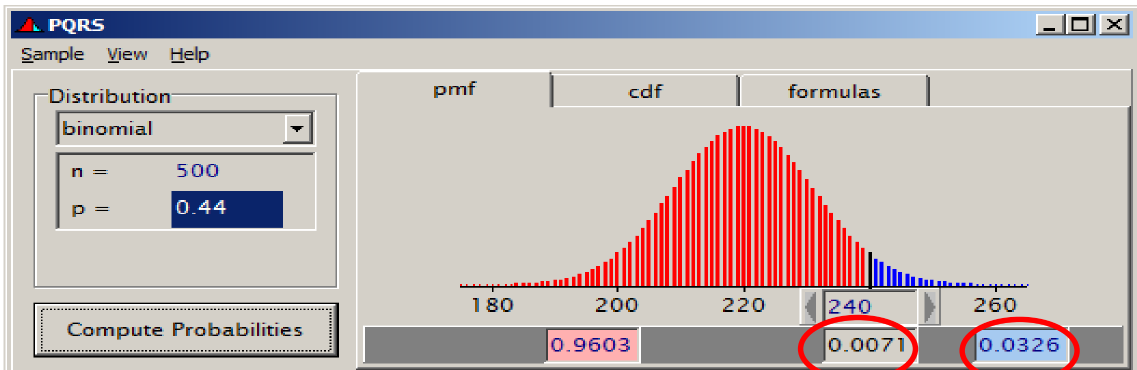
Before the party

1. $H_0 \colon \pi = 0.44 \quad H_a \colon \pi > 0.44$
2. Test statistic is the number of observed binge drinkers y
3. If H_0 is true, y ~ Binomial (500, 0.44)
4. Under H_a y tends to larger values.
5. So, we will use RPV.

During the party: Descriptive (Sample) Statistics

240 binge drinkers

The after party with PQRS



7. RPV = $P(y \geq 240) = P(y = 240) + P(y > 240)$
= 0.0071 + 0.0326 = 0.04.

8. RPV < 0.05, so H_0 is rejected, H_a is accepted. It is shown (α=0.05) that the proportion of binge drinkers at the university is larger than 0.44.

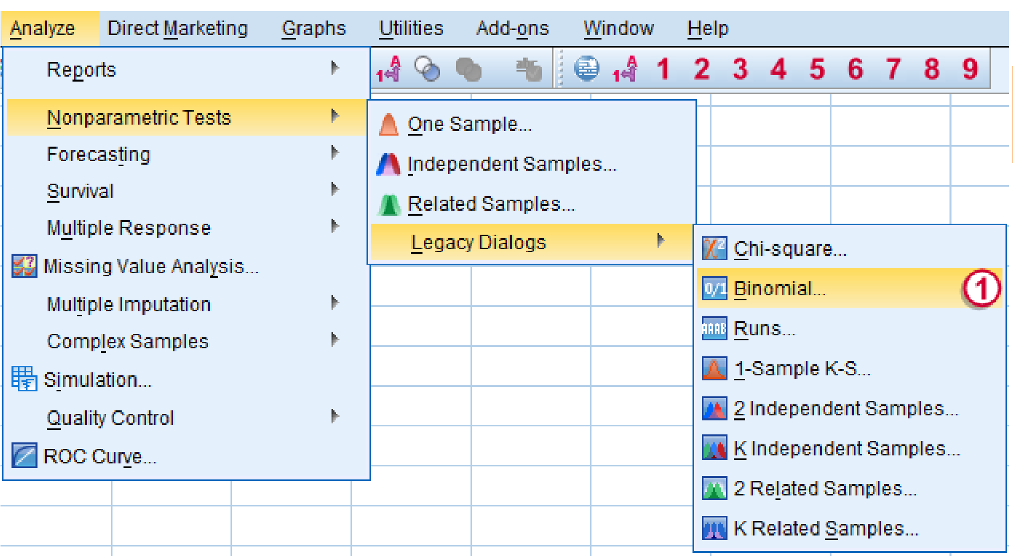
The after party with SPSS output

$\# \ of \ 'binge \ drinking \ students'$

$\hat{\pi}$

Binomial Test

		Category	N	Observed Prop.	Test Prop.	Exact Sig. (1-tailed)
Outcome	Group 1	Yes	240	.48	.44	.040
	Group 2	No	260	.52		
	Total		500	1.00		



$value \ under \ H_0$

$one \ tailed \ p - value$

The after party with R output

```
> binom.test(240,500,p=0.44,alternative = c("greater"),conf.level=0.95)

        Exact binomial test

data:  240 and 500
number of successes = 240, number of trials = 500, p-value = 0.03974
alternative hypothesis: true probability of success is greater than 0.44
95 percent confidence interval:
 0.4424098 1.0000000
sample estimates:
probability of success
                  0.48

> binom.test(240,500,p=0.44,alternative = c("two.sided"),conf.level=0.95)

        Exact binomial test

data:  240 and 500
number of successes = 240, number of trials = 500, p-value = 0.07884
alternative hypothesis: true probability of success is not equal to 0.44
95 percent confidence interval:
 0.4354394 0.5247984
sample estimates:
probability of success
                  0.48
```

Situation 10: 1 probability / fraction

Data: 1 sample $x_1,\ldots, x_n$ with binary outcomes (1 or 0).

Parameter of interest:
(the population fraction of successes)

$$\pi \quad (= \mu_x)$$

Estimator

$$\hat{\pi} = y / n \quad (= \bar{x})$$

Standard error of the estimator (CI calculation)

$$se(\hat{\pi}) = \sqrt{\hat{\pi}(1 - \hat{\pi}) / n}$$

Relevant distribution:

$$\text{y} \sim \text{Binomial}(\text{n}, \pi)$$

WAGENINGEN UR
For quality of life

Confidence intervals and sample sizes for a proportion

1- α two-sided confidence interval for π: Same structure as before, but this time using a z-value in stead of a t-value: table 2, bottom line.

$$\hat{\pi} \pm z_{\alpha/2} \cdot \sqrt{\hat{\pi}(1-\hat{\pi})/n}$$

- sample size n: same as before, but replace σ^2 by $\pi(1-\pi)$ e.g. if the aim is to calculate a (1-α)-confidence interval.

$$n = \frac{(z_{\alpha/2})^2 \times \sigma_x^2}{E^2} = \frac{(z_{\alpha/2})^2 \times \pi(1-\pi)}{E^2}$$

Use estimate for π, or largest value for $\pi(1-\pi) = 0.5 * 0.5 = 0.25$

WAGENINGEN UR
For quality of life

Alcohol abuse - confidence interval for π

estimate $\pm$ constant * standard error(estimate)

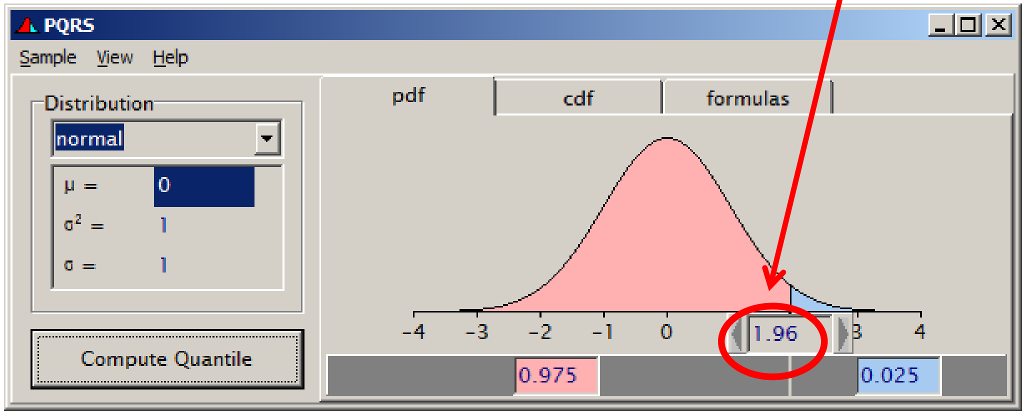
estimate:

$$\hat{\pi} = \frac{240}{500} = 0.48$$

$Z_{0.025}$ from N(0,1)
constant = 1.96

standard error (estimate):

$$\sqrt{\frac{0.48*(1-0.48)}{500}} = 0.0223$$



0.95-confidence interval:

$$(0.48 \pm 1.96 * 0.0223) =$$

$$(\mathbf{0.44}\ ,\ \mathbf{0.52})$$

WAGENINGEN UR
For quality of life

Alcohol abuse – sample size for width CI

The interval is (0.44, 0.52). Suppose we wanted a 0.95 CI with an expected width of 0.04, so E = 0.02.

New formula: $n = \frac{\left(z_{\alpha/2}\right)^2 \pi\,(1-\pi)}{E^2}$

when you do not have an estimate, use $\pi = 0.5 \rightarrow \pi\,(1-\pi)$=0.25

0.95 CI, so α = 0.05, so $z_{\alpha/2} = z_{0.025} = 1.96$

So: $n = \frac{1.96^2 * 0.2496}{0.02^2} = 2397.2$ rounded to n = 2398

$\pi(1-\pi) = \sigma^2 =$ variance for a binary observation.
In this case we could use $0.48 * 0.52 = 0.2496$

Consequence of too much statistics: alcohol abuse ?!

Break!!!

Let's go drink something

to recover !!!

Situation 11. Food additive example, 2 proportions

	Like	Do not like	Total
Without additive	5	9	14
With additive	12	4	16
Total	17	13	30

Question. Do consumers like the taste of a particular product more or less when a particular additive is used?

Setup: CRD with t=2. From a population, 30 people are randomly assigned to portions of food without or with the additive.
Each person is asked whether he or she liked the taste of the product offered or not.

Setup. Due to an error there are 14 and 16 people in the groups.
Results. See table.

Situation 11. Food additive example, 2 proportions

We compare two proportions, say π_1 and π_2, from two 'populations'.

Population 1 (2) is the population of all consumers tasting the product without (with) the additive.

π_1 (π_2) is the population proportion of people that like the taste of the product without (with) the additive.

The experimental units are consumers.

Response x: consumer likes the product (x=1) or not (x=0). So it is a binary response. → **Situation 11**.

Situation 11. Comparing two proportions

Two independentent samples, with sizes n_1 and n_2, N=n_1+n_2

Success probabilities π_1 and π_2,

Parameter of interest: $\boldsymbol{\pi_1 - \pi_2}$

y_1, y_2: Number of sample successes

Estimator: $\hat{\pi}_1 - \hat{\pi}_2 = \frac{y_1}{n_1} - \frac{y_2}{n_2}$

se($\hat{\pi}_1$- $\hat{\pi}_2$): $\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}$

$(\mathit{for\ CI - calculation})$

approximate (1-α) C.I.:

$$\hat{\pi}_1 - \hat{\pi}_2 \pm z_{\alpha/2} \cdot se(\hat{\pi}_1 - \hat{\pi}_2)$$

check for both samples
$n\,\pi$ and $n\,(1 - \pi) > 5$

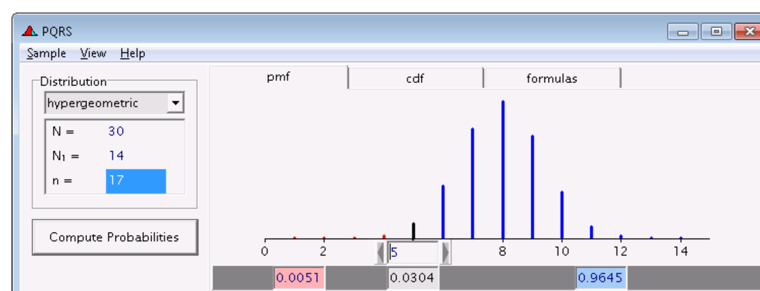
Fisher exact test

Relevant distribution for the Test Statistic:

Hypergeometric (N, n_1, y_1+y_2)

Before the party

1. $H_0: \pi_1 - \pi_2 = 0$,
   $H_a: \pi_1 - \pi_2 \neq 0$
2. Test statistic is y - the number of Likes in the "without additive" sample
3. If H_0 is true, y ~ HyperG (30, 14, 17)
4. Under H_a y tends to smaller or larger values.
5. So, we will use 2 - PV.



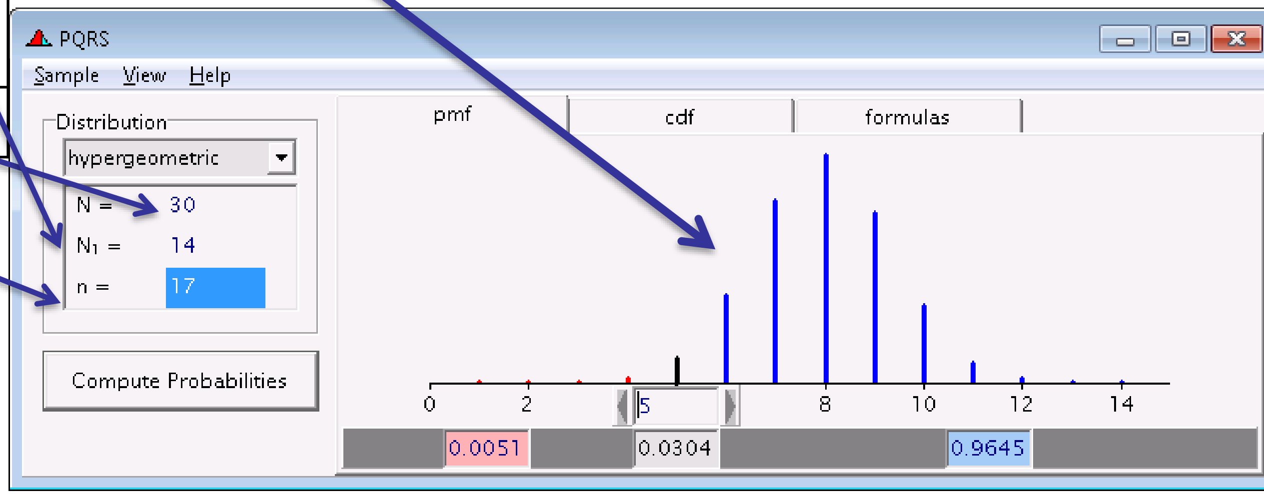
During the party: Descriptive (Sample) Statistics

	Like	do not like	Total
without	5	9	14
with	12	4	16
Total	17	13	30

The after party with PQRS

	Like	do not like	Total
without	5	9	14
with	12	4	16
Total	17	13	30

Not a symmetric distribution!!



7. LPV = 0.0051+0.0304= 0.0355

2tailed PV = 0.071 (2x LPV)

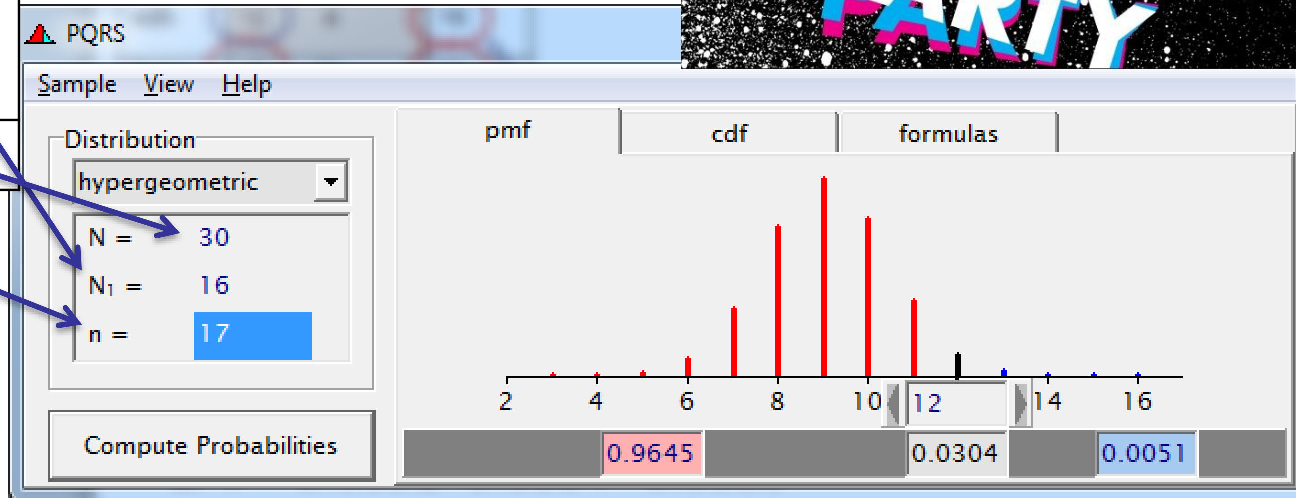
Note: for non-symmetric distributions there are other ways to calculate the 2tailed PV. We will not discuss that.

8. 0.071 > 0.05. So H_0 is not rejected, H_a is not accepted. It is not shown (α=0.05) that the additive affects the proportion of consumers that likes the taste of the product.

24

The after party with PQRS (other choice of TS)

	Like	do not like	Total
without	5	9	14
with	12	4	16
Total	17	13	30



P(X<5) P(X=5) P(X>5)

Suppose we had chosen: y_2 = the number of Likes for the product with additive. Then under H_0, given the margninal totals: $y_2 \sim$ HyperG(30, 16, 17) or HyperG(30, 17, 16).

Now: RPV = 0.0051+0.0304= 0.0355

2tailed PV = 0.071 (2x RPV).

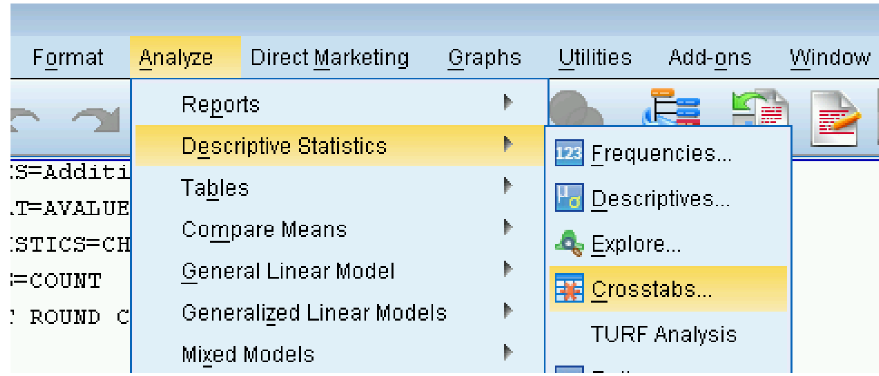
Result will be the same.

The after party with SPSS

Additive * Like Crosstabulation

Count

		Like		Total
		Yes	No	
Additive	Without	5	9	14
	With	12	4	16
Total		17	13	30

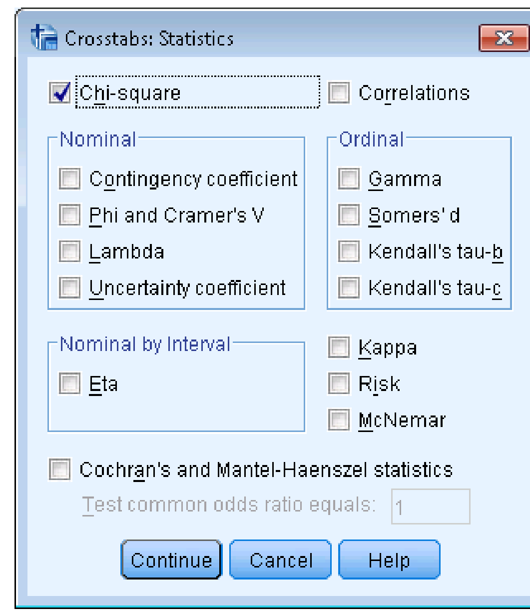


Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	4.693[a]	1	.030		
Continuity Correction[b]	3.229	1	.072		
Likelihood Ratio	4.810	1	.028		
Fisher's Exact Test				.063	.035
Linear-by-Linear Association	4.537	1	.033		
N of Valid Cases	30				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 6.07.

b. Computed only for a 2x2 table



NB: The 2 tailed P-value in SPSS is here not 2x 1 tailed P-value

0.95 CI for difference two proportions – food additive

	Like	Do not like	Total
Without additive	5	9	14
With additive	12	4	16
Total	17	13	30

An approximate two-sided 0.95 confidence interval for $\pi_1 - \pi_2$ is given by

$$\boxed{\hat{\pi}_1 - \hat{\pi}_2 \pm z_{\alpha/2} \cdot se(\hat{\pi}_1 - \hat{\pi}_2)}$$

$\hat{\pi}_1 - \hat{\pi}_2$ = 5/14 – 12/16 = - 0.392

$se(\hat{\pi}_1 - \hat{\pi}_2) = \sqrt{\frac{\frac{5}{14}*\frac{9}{14}}{14} + \frac{\frac{12}{16}*\frac{4}{16}}{16}} = 0.16$

$Z_{0.025} = 1.96$ → EM = 1.96 x 0.167 = 0.328

So the 0.95 CI limits are: -0.392 ± 0.328 → .95 CI: (-0.719, -0.063)

Bonus ... for being so quiet ;)

Situation 12. 1 sample, 1 nominal variable

Test of a new drug applied to n = 200 patients that are classified in four classes for blood pressure (1, 2, 3, 4 = marked decrease, moderate decrease, slight decrease, stationary / slight increase of blood pressure).

Are proportions in the classes comparable to "known" proportions (0.50, 0.25, 0.10 and 0.15) of the standard therapy?

Ex.10.10 high blood pressure O&L

Consider a variable with K possible outcomes, often not ordinal. This is then a ***nominal*** variable. We call the probabilities for the outcomes: $\pi_1, \pi_2, \ldots \pi_K$, where $\pi_1 + \pi_2 + \ldots + \pi_K = 1$.

In an experiment with n observations, the frequencies of the K outcomes are called e.g. $n_1 \ldots n_K$, where $n_1 + \ldots + n_K = n$.

Note: $n_1 \ldots n_K$ are random variables. Together the vector $(n_1 \ldots n_K)$ has the so-called multinomial$(\pi_1, \pi_2, \ldots \pi_K)$ distribution. K=2→ binomial

Before the party

H_0: π_1=0.50, π_2=0.25, π_3=0.10, π_4=0.15
H_a: at least one π_i is not equal to the proportions above

$$\chi^2 = \sum_{i=1}^{K} \frac{(n_i - E_i)^2}{E_i}$$

where the $E_i = n\,\pi_{i0}$ are the expected counts under H_0 and K=4

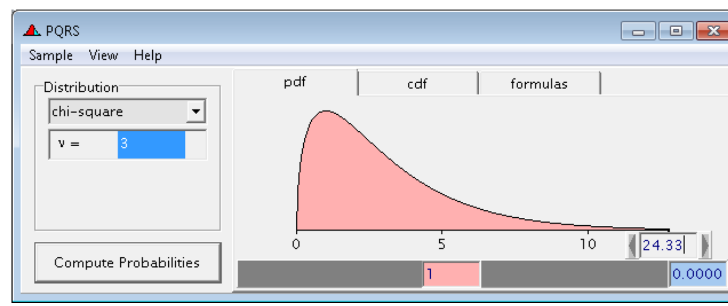
$$\chi^2 \overset{H_0}{\underset{approx.}{\sim}} \chi^2_{K-1}$$

Approximation is adequate if 100% of $E_i \geq 1$ and 80% of $E_i \geq 5$

It is also possible to use exact test.

Under H_a x^2 tends to larger values than under H_0

We use a right-sided R.R. / p value



$$H_0 : \pi_1 = \pi_{10} \ldots \pi_K = \pi_{K0}, \quad H_a : \pi_i \neq \pi_{i0}, \text{ for some } i = 1 \ldots K$$

During the party: Descriptive (Sample) Statistics

n_1 = 120, n_2 = 60,

n_3 = 10, n_4 = 10.

The after party with PQRS

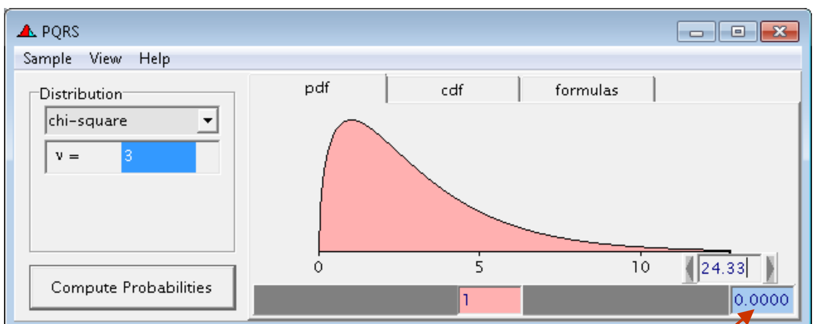
Outcome of the TS

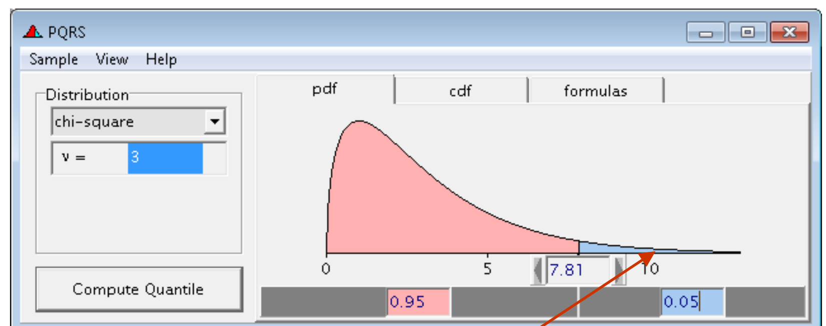
$n_1 = 120$, $n_2 = 60$, $n_3 = 10$, $n_4 = 10$.

$\chi^2 = 24.33$

for example: $E_1 = 200 * 0.5 = 100$

$E_1 = 100$, $E_2 = 50$, $E_3 = 20$, $E_4 = 30$.





P-value

Rejection Region (RR)

Table 7, O&L

H_0: $\pi_1 = 0.5$, $\pi_2 = 0.25$, $\pi_3 = 0.10$, $\pi_4 = 0.15$ is rejected: the new-drug effects are not comparable to the standard-drug effects

The after party with R / SPSS output

```
        Chi-squared test for given probabilities

data:  table(pressure)
X-squared = 24.333, df = 3, p-value = 2.128e-05
```

CATEGORY

	Observed N	Expected N	Residual
1	120	100.0	20.0
2	60	50.0	10.0
3	10	20.0	-10.0
4	10	30.0	-20.0
Total	200		

for example: $E_1 = 200 * 0.5 = 100$

$n_1 = 120$, $n_2 = 60$, … etc.

$E_1 = 100$, $E_2 = 50$, … etc.

Test Statistics

	CATEGORY
Chi-Square[a]	24.333
df	3
Asymp. Sig.	.000
Exact Sig.	.000
Point Probability	.000

a. 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 20.0.

$\chi^2 = 24.33$ with df = $K - 1 = 4 - 1 = 3$,

P-value = $P(\chi^2 > 24.33) = 0.000 < 0.05$

H_0: $\pi_1 = 0.5$, $\pi_2 = 0.25$, $\pi_3 = 0.10$, $\pi_4 = 0.15$ is rejected: the new-drug effects are not comparable to the standard-drug effects