CENTRE DE RECHERCHE PUBLIC

Microarray Center

the Luxembourg Bioinformatics Network

UNIVERSITÉ DU LUXEMBOURG

STATISTICAL DATA ANALYSIS IN EXCEL

Lecture 6

Some Advanced Topics

Dr. Petr Nazarov

petr.nazarov@crp-sante.lu

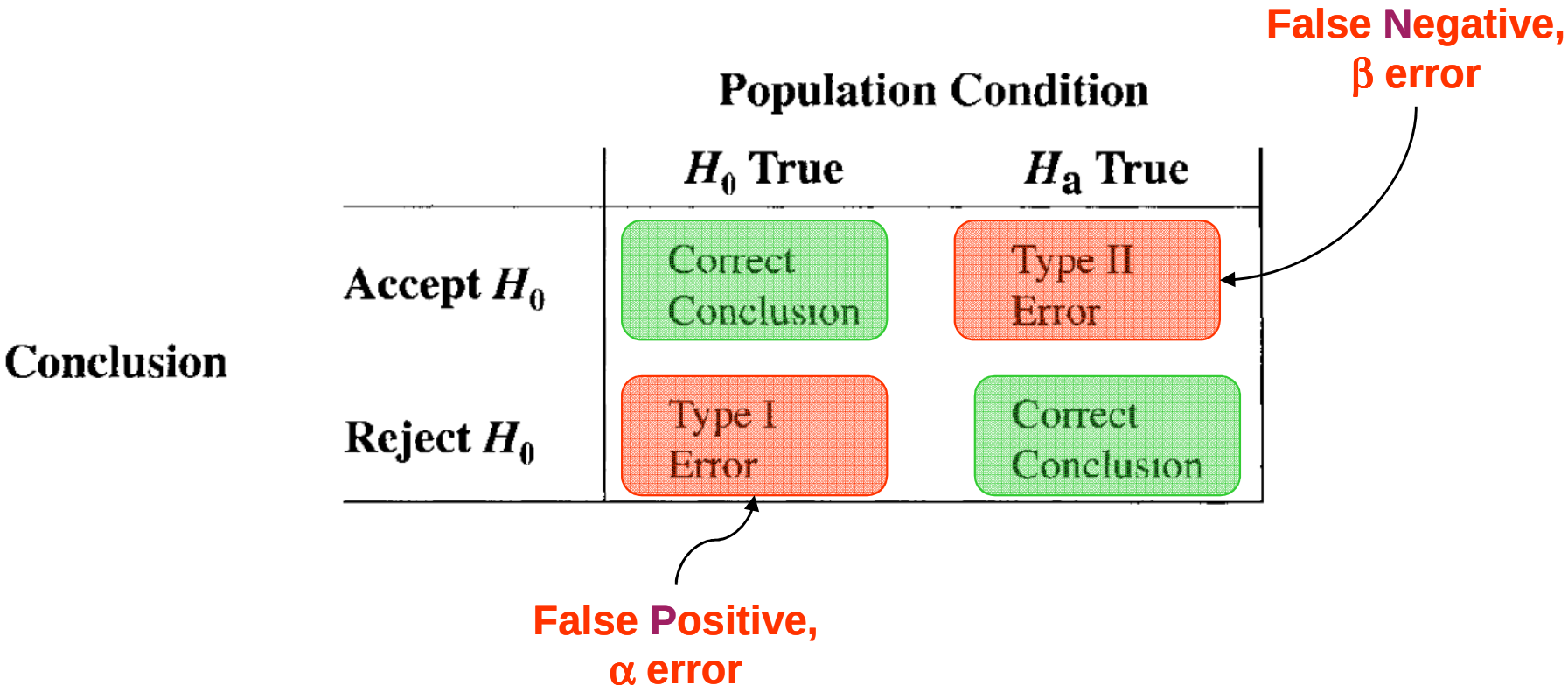
31-10-2011

Correction for Multiple Comparisons

Please download the data from edu.sablab.net/data/xls

`all_data.xls`

Correct Results and Errors



Probability of an error in a multiple test:

$1-(0.95)^{\text{number of comparisons}}$

False discovery rate (FDR)

FDR control is a statistical method used in multiple hypothesis testing to correct for multiple comparisons. In a list of rejected hypotheses, FDR controls the expected proportion of incorrectly rejected null hypotheses (type I errors).

Conclusion \ Population Condition	H_0 is TRUE	H_0 is FALSE	Total
Accept H_0 (non-significant)	U	T	$m - R$
Reject H_0 (significant)	V	S	R
Total	m_0	$m - m_0$	m

$$FDR = E\left(\frac{V}{V+S}\right)$$

Assume we need to perform *k* = 100 comparisons, and select maximum FDR = α = 0.05

Independent tests

The Simes procedure ensures that its expected value $\mathrm{E}\left[\frac{V}{V+S}\right]$ is less than a given α (Benjamini and Hochberg 1995). This procedure is valid when the m tests are independent. Let $H_1 \ldots H_m$ be the null hypotheses and $P_1 \ldots P_m$ their corresponding p-values. Order these values in increasing order and denote them by $P_{(1)} \ldots P_{(m)}$. For a given α, find the largest k such that $P_{(k)} \leq \frac{k}{m}\alpha.$

Then reject (i.e. declare positive) all $H_{(i)}$ for $i = 1, \ldots, k.$

Note that the mean α for these m tests is $\frac{\alpha(m+1)}{2m}$ which could be used as a rough FDR, or RFDR, "α adjusted for m indep. tests." The RFDR calculation shown here provides a useful approximation and is not part of the Benjamini and Hochberg method; see AFDR below.

False Discovery Rate

Assume we need to perform k = 100 comparisons, and select maximum FDR = α = 0.05

$$FDR = E\left(\frac{V}{V+S}\right)$$

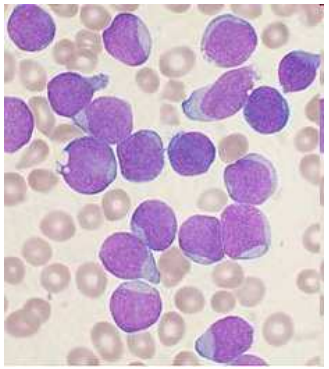
Expected value for FDR < α if $P_{(k)} \leq \frac{k}{m}\alpha$

$$\frac{mP_{(k)}}{k} \leq \alpha$$

Example: Acute Lymphoblastic Leukemia

all_data.xls

Acute lymphoblastic leukemia (ALL), is a form of leukemia, or cancer of the white blood cells characterized by excess lymphoblasts.



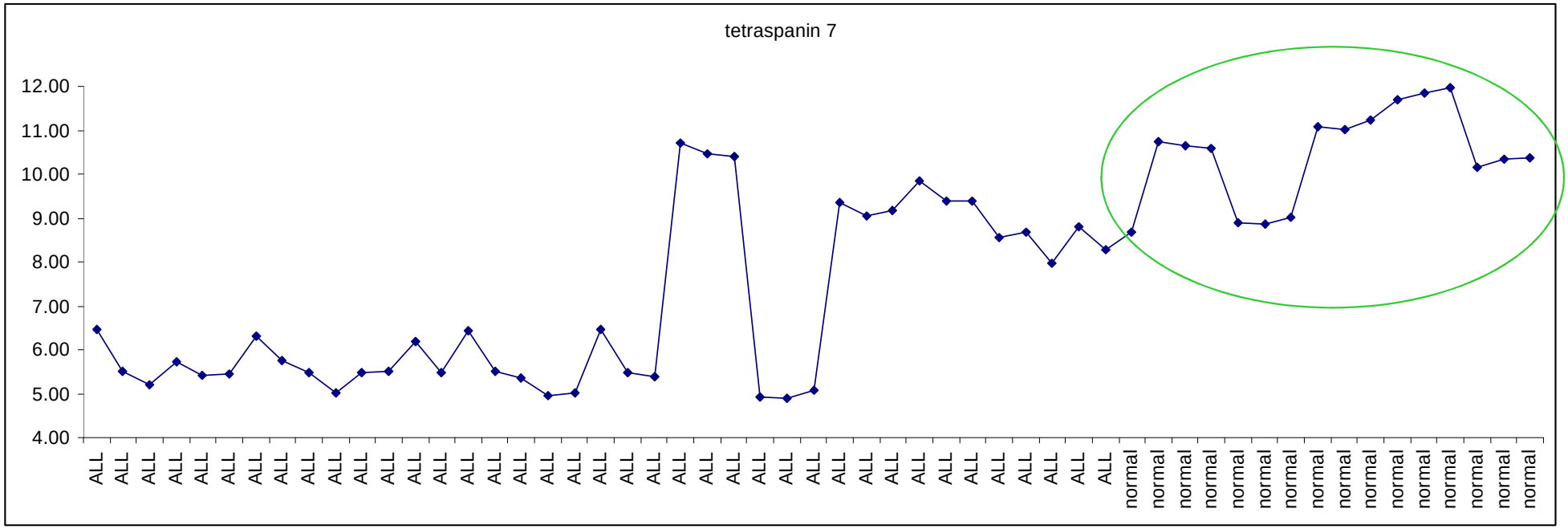
all_data.xls contains the results of full-trancript profiling for ALL patients and healthy donors using Affymetrix microarrays. The data were downloaded from ArrayExpress repository and normalized. The expression values in the table are in $\log_2$ scale.

Let us analyze these data:

- Calculate log-ratio (logFC) for each gene
- Calculate the p-value based on t-test for each gene
- Perform the FDR-based adjustment of the p-value.

  Calculate the number of up and down regulated genes with FDR<0.01
- How would you take into account logFC?

Example score: $score = -\log(adj.p.value) \cdot |logFC|$



look for "tetraspanin 7" + leukemia in google ☺

Results are never perfect…

Empirical Interval Estimation for Random Functions

INTERVAL ESTIMATIONS FOR RANDOM FUNCTIONS

Sum and Square of Normal Variables

Distribution of sum or difference of 2 normal random variables

The sum/difference of 2 (or more) normal random variables is a normal random variable with mean equal to sum/difference of the means and variance equal to **SUM** of the variances of the compounds.

$$x \pm y \rightarrow Normal\ distribution$$

$$E[x \pm y] = E[x] \pm E[y]$$

$$\sigma^2_{x \pm y} = \sigma^2_x + \sigma^2_y$$

Distribution of sum of squares on *k* standard normal random variables

The sum of squares of *k* standard normal random variables is a χ^2 with *k* degree of freedom.

$$if \quad x_1, \ldots, x_k \rightarrow Normal\ distribution$$

$$\sum_{i=1}^{k} x_i^2 \rightarrow \chi^2 \quad with\ d.f. = k$$

What to do in more complex situations?

$$\frac{x}{y} \rightarrow ?$$

$$\sqrt{x} \rightarrow ?$$

$$\log(|x|) \rightarrow ?$$

Try to solve analytically?

Simplest case. E[x] = E[y] = 0

Ratio distribution

From Wikipedia, the free encyclopedia

A **ratio distribution** (or *quotient distribution*) is a probability distribution constructed as the distribution of the ratio of random variables having two other known distributions. Given two random variables X and Y, the distribution of the random variable Z that is formed as the ratio

$$Z = X/Y$$

is a *ratio distribution*.

$$p_Z(z) = \frac{b(z)\cdot c(z)}{a^3(z)}\frac{1}{\sqrt{2\pi}\sigma_x\sigma_y}\left[2\Phi\left(\frac{b(z)}{a(z)}\right)-1\right]+\frac{1}{a^2(z)\cdot\pi\sigma_x\sigma_y}e^{-\frac{1}{2}\left(\frac{\mu_x^2}{\sigma_x^2}+\frac{\mu_y^2}{\sigma_y^2}\right)}$$

where

$$a(z) = \sqrt{\frac{1}{\sigma_x^2}z^2+\frac{1}{\sigma_y^2}}$$

$$b(z) = \frac{\mu_x}{\sigma_x^2}z+\frac{\mu_y}{\sigma_y^2}$$

$$c(z) = e^{\frac{1}{2}\frac{b^2(z)}{a^2(z)}-\frac{1}{2}\left(\frac{\mu_x^2}{\sigma_x^2}+\frac{\mu_y^2}{\sigma_y^2}\right)}$$

$$\Phi(z) = \int_{-\infty}^{z}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}u^2}\,du$$

INTERVAL ESTIMATIONS FOR RANDOM FUNCTIONS

Practical Approach

Two rates where measured for a PCR experiment: experimental value (X) and control (Y). 5 replicates where performed for each.

From previous experience we know that the error between replicates is normally distributed.

Q1: provide an interval estimation for the fold change X/Y (α=0.05)

Q2: provide an interval estimation for the log fold change $\log_2(X/Y)$

#	Experiment	Control
1	215	83
2	253	75
3	198	62
4	225	91
5	240	70

Mean	226.2	76.2
StDev	21.39	11.26

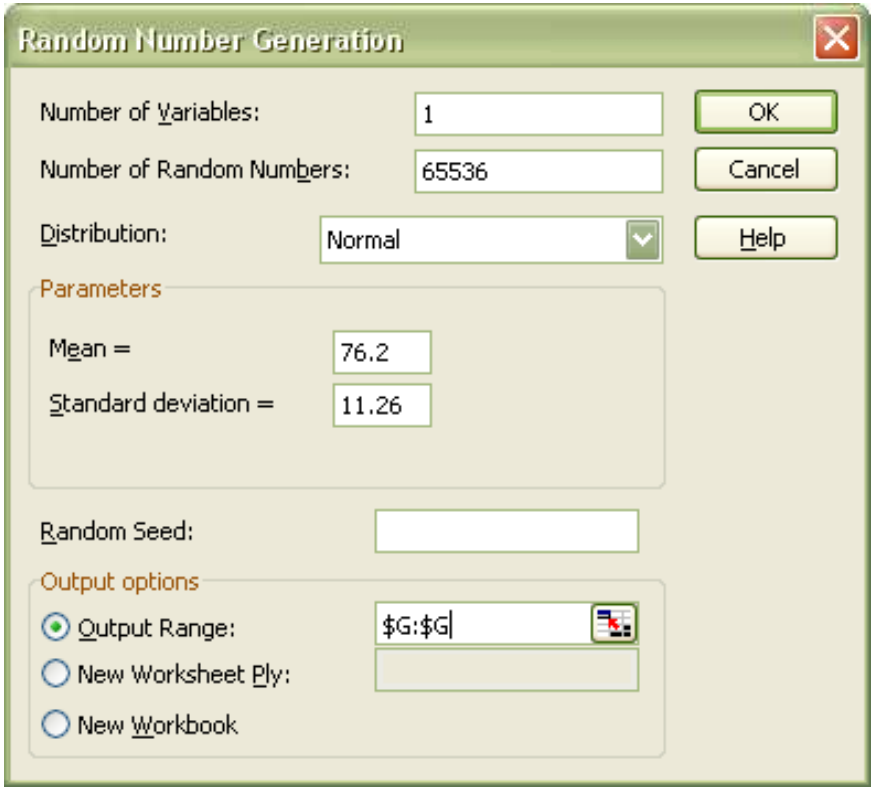
Let us use a ***numerical simulation…***

1. Generate 2 sets of 65536 normal random variable with means and standard deviations corresponding to ones of experimental and control set.

Mean	226.2	76.2
StDev	21.39	11.26

In Excel go: Tools → Data Analysis:

- **`Random Number Generation`**



If you do not have Data Analysis tool – approximate normal distribution by sum of uniform:

$$N(x, m_x, \sigma_x) = m_x + \sigma_x \left(\sum_{i=1}^{12} U(x_i) - 6 \right)$$

- **`= RAND()`** ← $U(x)$

INTERVAL ESTIMATIONS FOR RANDOM FUNCTIONS

Practical Approach

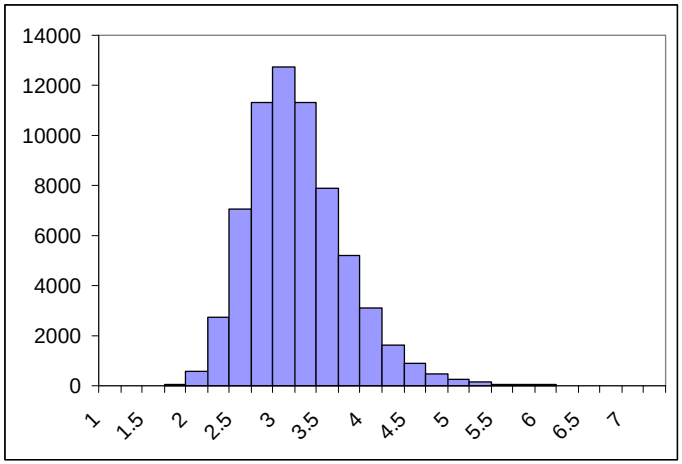
1. Generate 2 sets of 65536 normal random variable with means and standard deviations corresponding to ones of experimental and control set.

Mean	226.2	76.2
StDev	21.39	11.26
sim.m	226.088799	76.2823
sim.s	21.379652	11.2885

2. Build the target function. For Q1 build X/Y

X/Y.m	3.03289298
X/Y.s	0.566865
min	-8.14098141
max	7.72162205

3. Study the target function. Calculate summary, build histogram.



4. If you would like to have 95% interval, calculate 2.5% and 97.5% percentiles.

In Excel use function

- `=PERCENTILE(data,0.025)`

$X/Y \in [\ 2.13,\ 4.33\ ]$

INTERVAL ESTIMATIONS FOR RANDOM FUNCTIONS

Practical Approach

What was a "mistake" in the previous case?

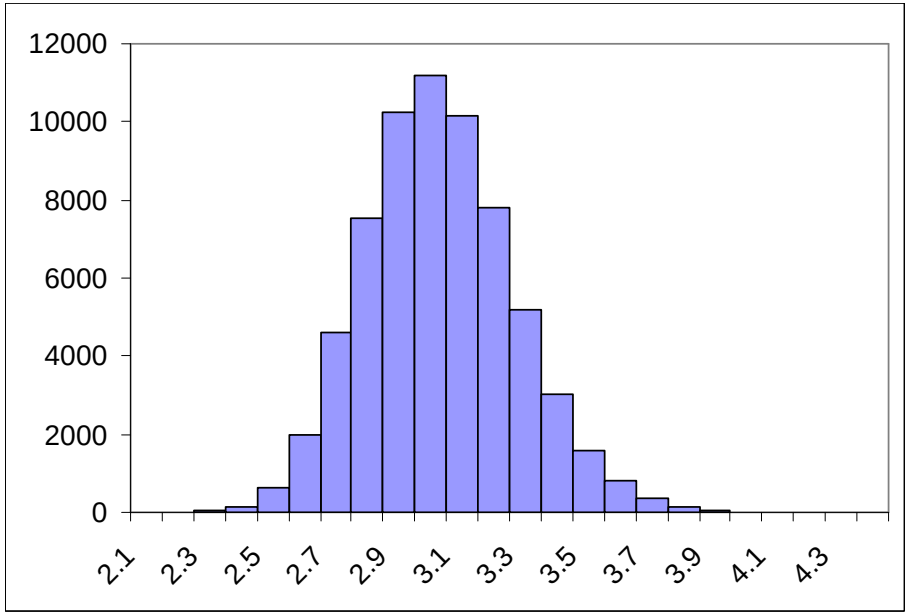
$$\sigma_m = \frac{\sigma}{\sqrt{n}}$$

There we spoke about **prediction interval** of X/Y. Now let's produce the **interval estimation for mean X/Y**

Mean	226.2	76.2
StDev	**9.57**	**5.03**

X/Y.m	2.98047943
X/Y.s	0.23616818
min	2.01556098
max	4.31131109

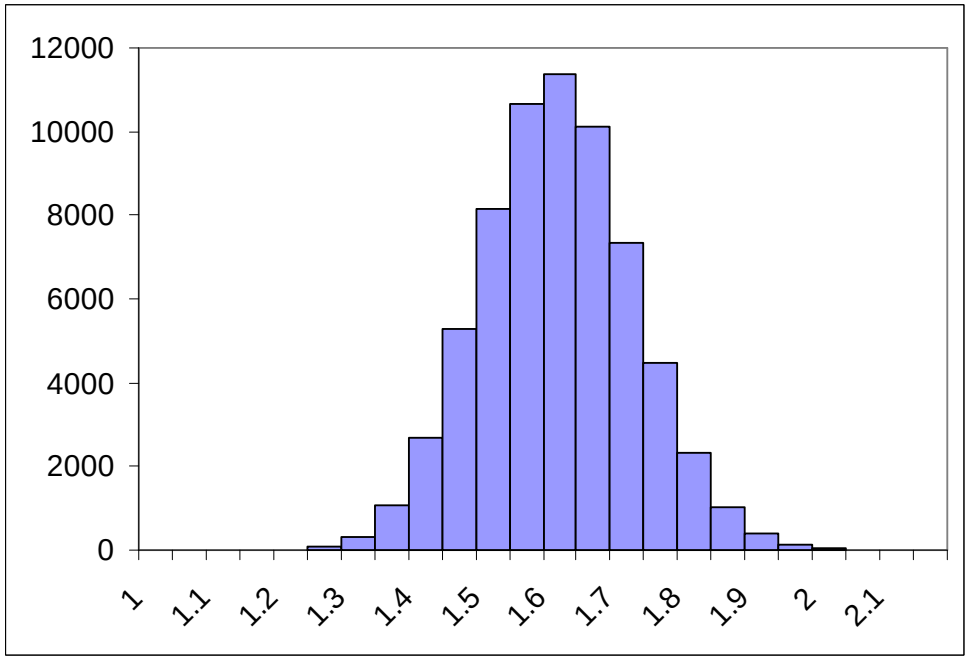
$E[X/Y] \in [\ 2.55,\ 3.48\ ]$



Q2: provide an interval estimation for the log fold change log2(X/Y)

Mean	1.571052
Standard Dev	0.113705

$E[\log(X/Y)] \in [\ 1.35,\ 1.80\ ]$



	Simulation	Normal
2.50%	1.3546	1.3482
97.50%	1.7998	1.7939

Goodness of Fit and Independence

Multinomial population
A population in which each element is assigned to one and only one of several categories. The multinomial distribution extends the binomial distribution from two to three or more outcomes.

Contingency table = Crosstabulation
Contingency tables or crosstabulations are used to record, summarize and analyze the relationship between two or more categorical (usually) variables.

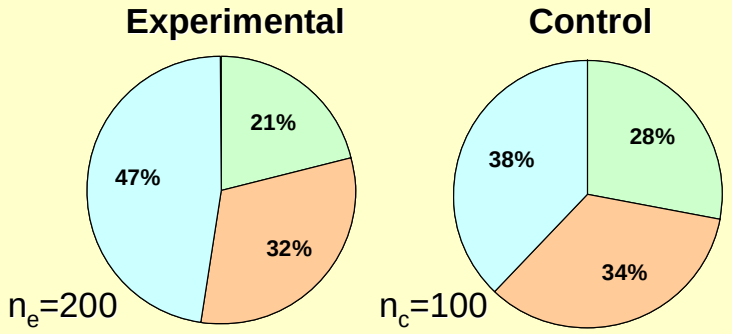
The new treatment for a disease is tested on 200 patients. The outcomes are classified as:

- **A** – patient is **completely treated**
- **B** – disease transforms into a **chronic form**
- **C** – treatment is **unsuccessful** ☹

In parallel the 100 patients treated with standard methods are observed

Category	Experimental	Control
A	94	38
B	42	28
C	64	34
Sum	200	100

◆ The proportions for 3 "classes" of patients with and without treatment are:



Are the proportions ***significantly different*** in control and experimental groups?

TEST OF GOODNESS OF FIT

Goodness of Fit

Goodness of fit test
A statistical test conducted to determine whether to reject a hypothesized probability distribution for a population.

Model – our assumption concerning the distribution, which we would like to test.

Observed frequency – frequency distribution for experimentally observed data, f_i

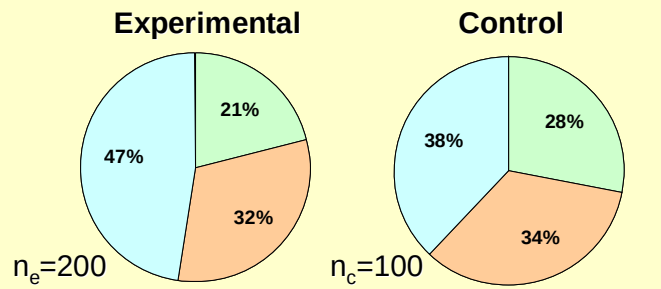
Expected frequency – frequency distribution, which we would expect from our **model**, e_i

Hypotheses for the test:

H_0: the population follows a multinomial distribution with the probabilities, specified by **model**

H_a: the population does not follow … **model**

The proportions for 3 "classes" of patients with and without treatment are:



Are the proportions ***significantly different*** in control and experimental groups?

Test statistics for goodness of fit

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i}$$

χ^2 has $k-1$ degree of freedom

At least 5 expected must be in each category!

TEST OF GOODNESS OF FIT

Example

The new treatment for a disease is tested on 200 patients. The outcomes are classified as:

- **A** – patient is **completely treated**
- **B** – disease transforms into a **chronic form**
- **C** – treatment is **unsuccessful** ☹

In parallel the 100 patients treated with standard methods are observed

Category	Experimental	Control
A	94	38
B	42	28
C	64	34
Sum	200	100

1. Select the model and calculate expected frequencies

Let's use control group (classical treatment) as a model, then:

2. Compare expected frequencies with the experimental ones and build χ^2

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i}$$

Category	Control frequencies	Model for control	Expected freq., e
A	38	0.38	76
B	28	0.28	56
C	34	0.34	68
Sum	100	1	**200**

Experimental freq., f
94
42
64
200

Category	**(f-e)2/e**
A	4.263
B	3.500
C	0.235
Chi2	**7.998**

3. Calculate p-value for χ^2 with d.f. = k–1

- = **CHIDIST(**χ^2,d.f.**)**
- = **CHITEST(**f, e**)**

p-value = 0.018, reject H_0

Goodness of Fit for Independence Test: Example

Alber's Brewery manufactures and distributes three types of beer: **white**, **regular**, and **dark**. In an analysis of the market segments for the three beers, the firm's market research group raised the question of whether preferences for the three beers differ among **male** and **female** beer drinkers. If beer preference is independent of the gender of the beer drinker, one advertising campaign will be initiated for all of Alber's beers. However, if beer preference depends on the gender of the beer drinker, the firm will tailor its promotions to different target markets.

beer.xls

H_0: Beer preference is **independent** of the gender of the beer drinker

H_a: Beer preference is **not independent** of the gender of the beer drinker

sex\beer	White	Regular	Dark	**Total**
Male	20	40	20	**80**
Female	30	30	10	**70**
Total	**50**	**70**	**30**	**150**

Goodness of Fit for Independence Test: Example

1. Build model assuming independence

sex\beer	White	Regular	Dark	Total
Male	20	40	20	**80**
Female	30	30	10	**70**
Total	**50**	**70**	**30**	**150**

	White	Regular	Dark	Total
Model	0.3333	0.4667	0.2000	1

2. Transfer the model into expected frequencies, multiplying model value by number in group

sex\beer	White	Regular	Dark	Total
Male	26.67	37.33	16.00	**80**
Female	23.33	32.67	14.00	**70**
Total	**50**	**70**	**30**	**150**

$$e_{ij} = \frac{(Row\ i\ Total)(Column\ \ j\ Total)}{Sample\ Size}$$

3. Build χ^2 statistics

$$\chi^2 = \sum_{i}^{n}\sum_{j}^{m}\frac{(f_{ij} - e_{ij})^2}{e_{ij}}$$

χ^2 =6.122

χ^2 distribution with d.f.=$(n-1)(m-1)$, provided that the expected frequencies are 5 or more for all categories.

4. Calculate p-value

p-value = 0.047, reject H_0

Test for Normality: Example

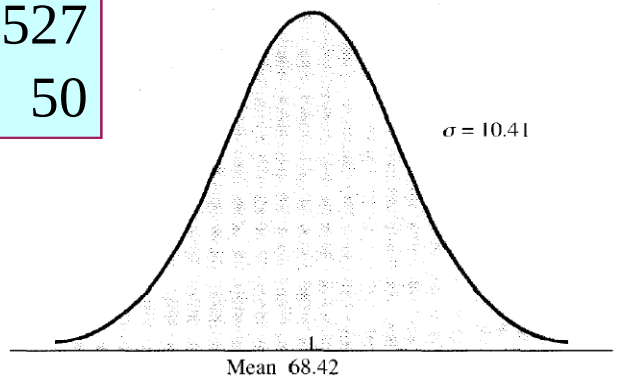
Chemline hires approximately 400 new employees annually for its four plants. The personnel director asks whether a normal distribution applies for the population of aptitude test scores. If such a distribution can be used, the distribution would be helpful in evaluating specific test scores; that is, scores in the upper 20%, lower 40%, and so on, could be identified quickly. Hence, we want to test the null hypothesis that the population of test scores has a normal distribution. The study will be based on 50 results.

chemline.xls

Aptitude test scores

71	86	56	61	65
60	63	76	69	56
55	79	56	74	93
82	80	90	80	73
85	62	64	54	54
65	54	63	73	58
77	56	65	76	64
61	84	70	53	79
79	61	62	61	65
66	70	68	76	71

Mean	68.42
Standard Deviation	10.4141
Sample Variance	108.4527
Count	50



H_0: The population of test scores **has a normal distribution** with mean 68.42 and standard deviation 10.41

H_a: the population **does not have** a mentioned distribution

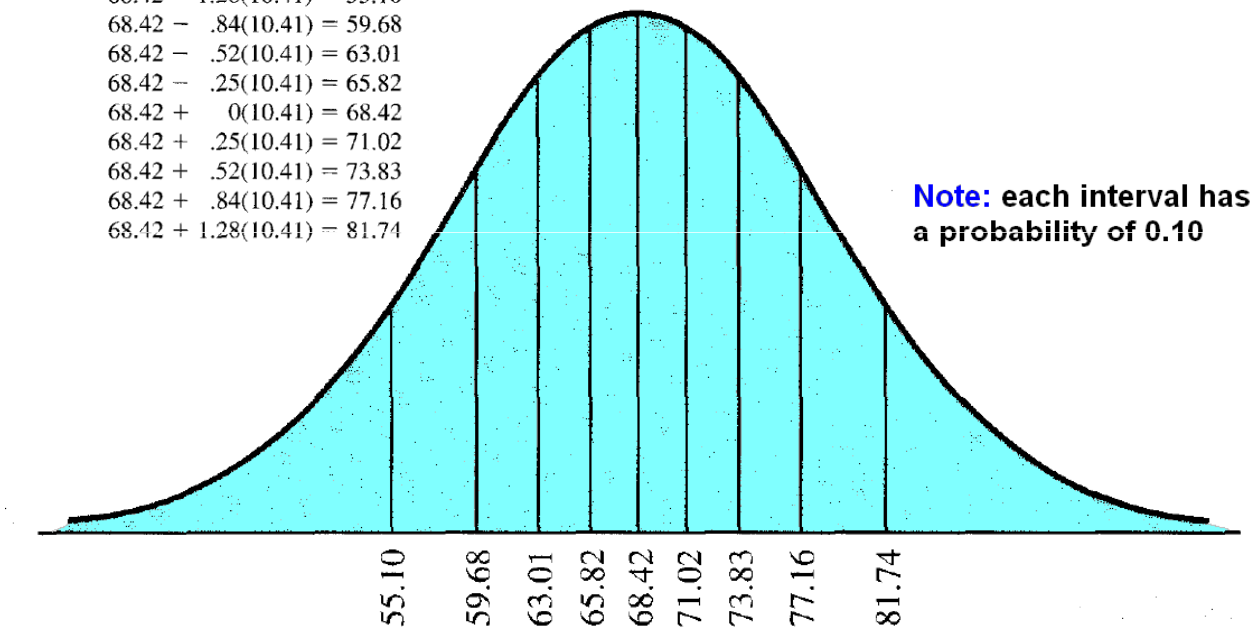
Test for Normality: Example

chemline.xls

Mean	68.42
Standard Deviation	10.4141
Sample Variance	108.4527
Count	50

Bin	*Observed frequency*	*Expected frequency*
55.1	5	5
59.68	5	5
63.01	9	5
65.82	6	5
68.42	2	5
71.02	5	5
73.83	2	5
77.16	5	5
81.74	5	5
More	6	5
Total	**50**	**50**

Lower 10%: 68.42 − 1.28(10.41) = 55.10
Lower 20%: 68.42 − .84(10.41) = 59.68
Lower 30%: 68.42 − .52(10.41) = 63.01
Lower 40%: 68.42 − .25(10.41) = 65.82
Mid-score: 68.42 + 0(10.41) = 68.42
Upper 40%: 68.42 + .25(10.41) = 71.02
Upper 30%: 68.42 + .52(10.41) = 73.83
Upper 20%: 68.42 + .84(10.41) = 77.16
Upper 10%: 68.42 + 1.28(10.41) = 81.74



Note: each interval has a probability of 0.10

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i}$$

χ^2 distribution with d.f.= $n - p - 1$,

where p – number of estimated parameters

$p = 2$ includes mean and variance
d.f. = 10 – 2 – 1
$\chi^2 = 7.2$

p-value = 0.41, cannot reject H_0

Thank you for your attention