4V001 – Workshops in molecular and cellular biology

Applied biostatistics

Martin LARSEN

www.immulab.fr/cms/index.php/teaching

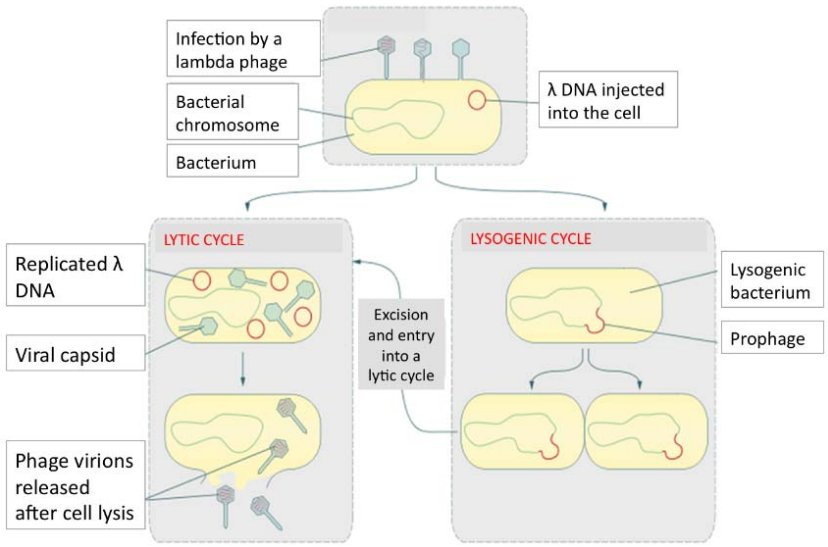
INSERM U1135, CHU Pitié-Salpetrière, Paris, France

Hypothesis testing Step-by-step

1. Define the Problem
2. State the Objectives
3. State the Null Hypothesis (H_0)
4. State the Alternative Hypothesis (H_1)
5. Select the appropriate statistical test
6. Decide if the Hypothesis testing will be left-tailed, right-tailed, or two tailed test.
7. State the alpha-risk (α) level
8. State the beta-risk (1-β) level
9. State or Establish (require prior knowledge) the Effect Size
10. Create Sampling Plan, determine sample size
11. Gather samples
12. Collect and pre-analyse data
13. Calculate the test statistic
14. Determine critical test value and p-value
    - If p-value < α, reject H_0
    - If p-value > α, fail to reject H_0
15. *Post hoc* analysis

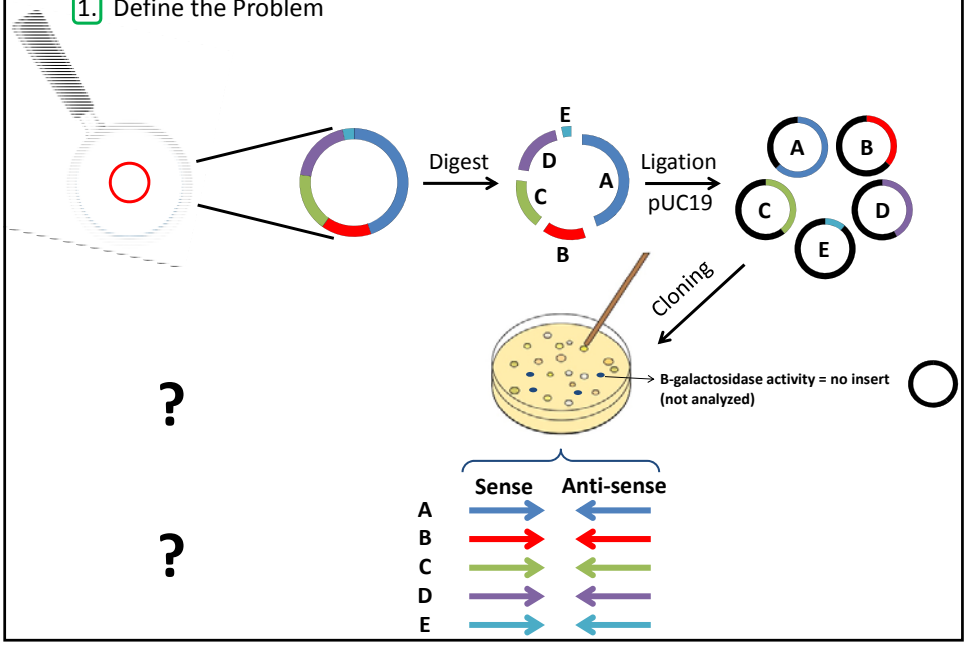
Hypothesis testing – What's the problem

1. Define the Problem



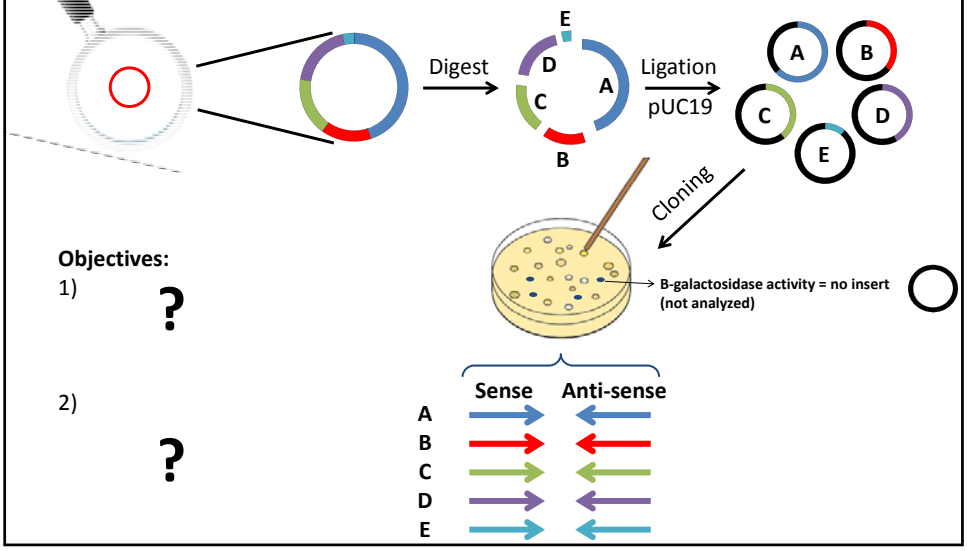
Hypothesis testing – What's the problem

1. Define the Problem



Hypothesis testing - Objectives

1. Define the Problem
2. State the Objectives



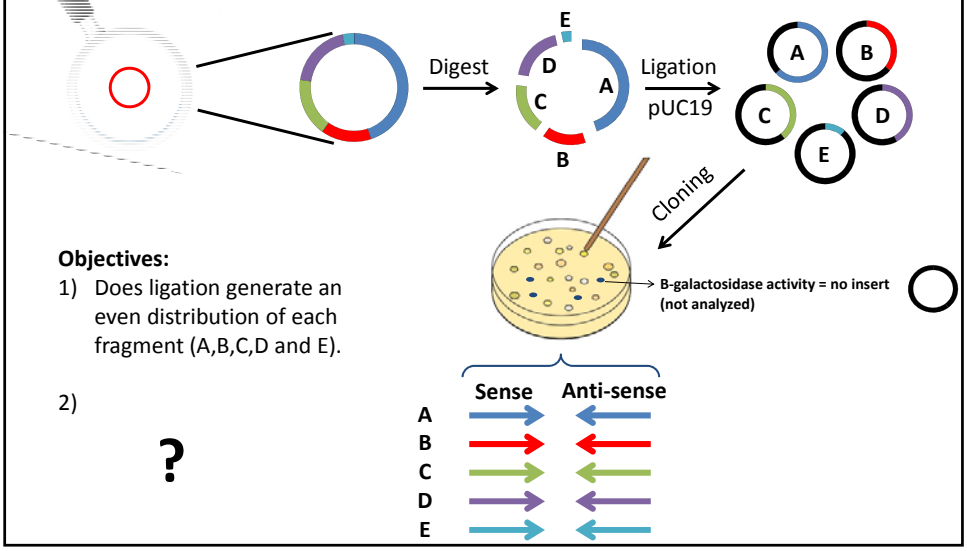
Objectives:

1) ?

2) ?

Hypothesis testing - Objectives

1. Define the Problem
2. State the Objectives



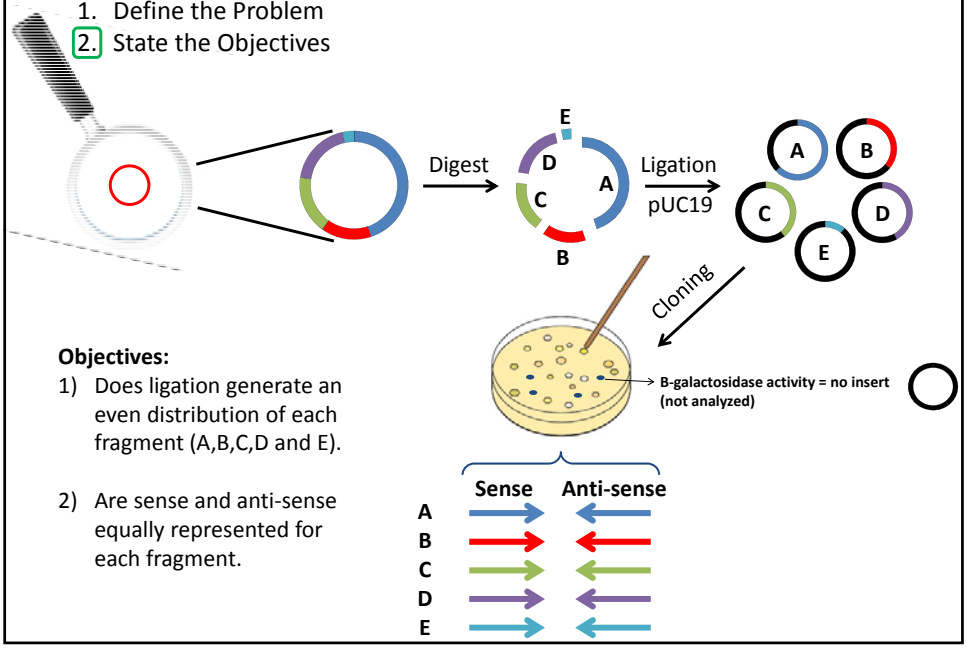
Objectives:

1) Does ligation generate an even distribution of each fragment (A,B,C,D and E).

2) ?

Hypothesis testing - Objectives

1. Define the Problem
2. State the Objectives



Objectives:

1) Does ligation generate an even distribution of each fragment (A,B,C,D and E).

2) Are sense and anti-sense equally represented for each fragment.

Hypothesis testing - State hypothesis

1. Define the Problem
2. State the Objectives
   - Does ligation generate an even distribution of each fragment (A,B,C,D and E).
   - Is sense and anti-sense equally represented for each fragment.
3. State the Null Hypothesis (H_0)
   - H_0: Fragments are distributed
     - Obj. 1: ?
     - Obj. 2: ?
4. State the Alternative Hypothesis (H_1)
   - H_1: Fragments are not distributed
     - Obj. 1: ?
     - Obj. 2: ?

Hypothesis testing – State hypothesis

1. Define the Problem
2. State the Objectives
   - Does ligation generate an even distribution of each fragment (A,B,C,D and E).
   - Is sense and anti-sense equally represented for each fragment.
3. State the Null Hypothesis (H_0)
   - H_0: Fragments are distributed
     - Obj. 1: uniformly
     - Obj. 2: ?
4. State the Alternative Hypothesis (H_1)
   - H_1: Fragments are not distributed
     - Obj. 1: uniformly
     - Obj. 2: ?

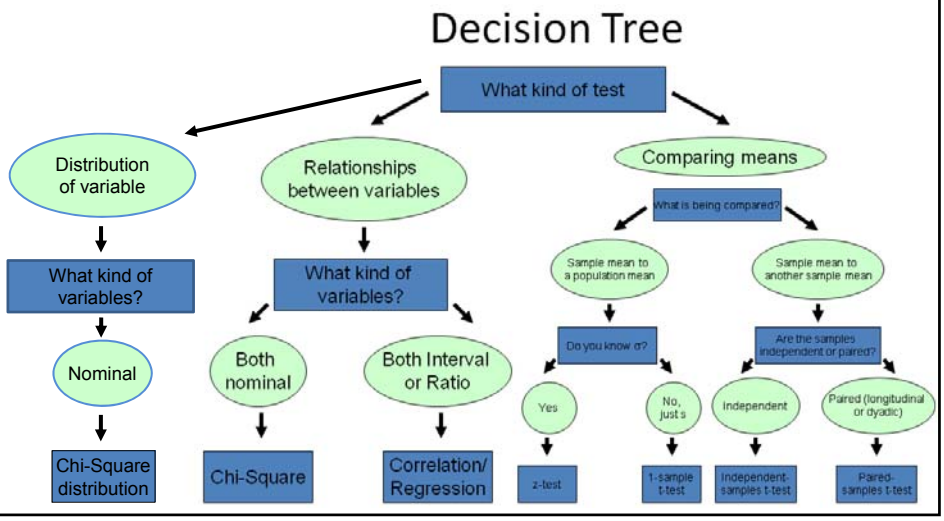
Hypothesis testing – State hypothesis

1. Define the Problem
2. State the Objectives
   - Does ligation generate an even distribution of each fragment (A,B,C,D and E).
   - Is sense and anti-sense equally represented for each fragment.
3. State the Null Hypothesis (H_0)
   - H_0: Fragments are distributed
     - Obj. 1: uniformly
     - Obj. 2: proportionally
4. State the Alternative Hypothesis (H_1)
   - H_1: Fragments are not distributed
     - Obj. 1: uniformly
     - Obj. 2: proportionally

Hypothesis testing – Which test?

- H_0: Fragments are distributed uniformly (O1) or proportionally (O2)
- H_1: Fragments are not distributed uniformly or proportionally

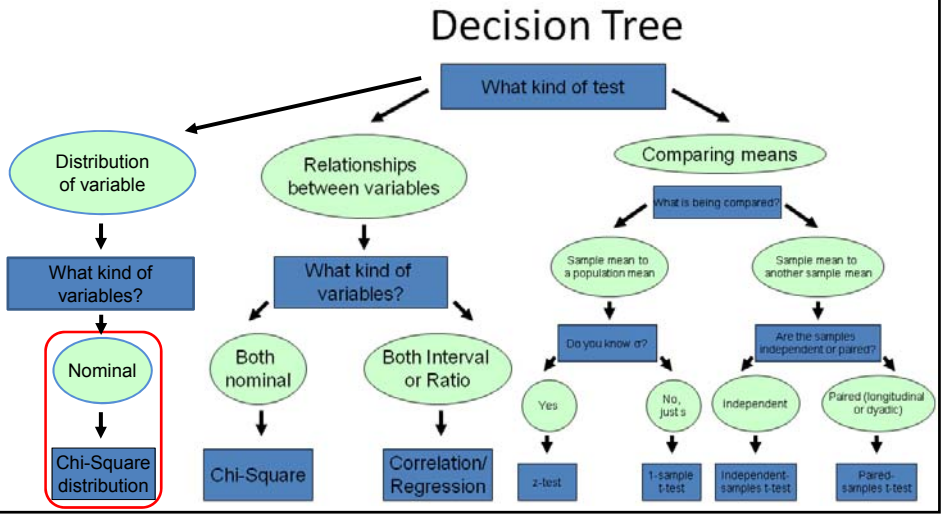
5. Select the appropriate statistical test



Hypothesis testing – Which test?

- H_0: Fragments are distributed uniformly (O1) or proportionally (O2)
- H_1: Fragments are not distributed uniformly or proportionally

5. Select the appropriate statistical test



Hypothesis testing – Which χ^2 test?

χ^2 - goodness-of fit

Measured variable

O_1	O_2	O_3	O_4

O_i is the i^{th} observed value

Expected values are known *a priori*

χ^2 - homogeneity

Measured variable

Group variable stratifying sample	O_1	O_2
	O_3	O_4

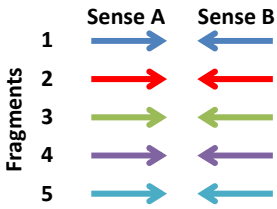
Expected values are defined based on H_0, most often column and row totals (proportional distribution)

χ^2 - independence

Measured variable 1

Measured variable 2	O_1	O_2
	O_3	O_4

Data structure



	A	B
1	O_{1A}	O_{1B}
2	O_{2A}	O_{2B}
3	O_{3A}	O_{3B}
4	O_{4A}	O_{4B}
5	O_{5A}	O_{5B}

Objectives:

1) Does ligation generate an even distribution of each fragment (A,B,C,D and E).

2) Are sense and anti-sense equally represented for each fragment.

Hypothesis testing – Which χ^2 test?

χ^2 - goodness-of fit

Measured variable

O_1	O_2	O_3	O_4

O_i is the i^{th} observed value

Expected values are known *a priori*

χ^2 - homogeneity

Measured variable

Group variable stratifying sample	O_1	O_2
	O_3	O_4

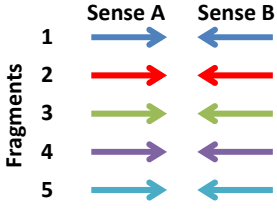
Expected values are defined based on H_0, most often column and row totals (proportional distribution)

χ^2 - independence

Measured variable 1

Measured variable 2	O_1	O_2
	O_3	O_4

Data structure



	A+B
1	$O_{1A}+O_{1B}$
2	$O_{2A}+O_{2B}$
3	$O_{3A}+O_{3B}$
4	$O_{4A}+O_{4B}$
5	$O_{5A}+O_{5B}$

Objectives:

1) Does ligation generate an even distribution of each fragment (A,B,C,D and E).

2) Are sense and anti-sense equally represented for each fragment.

Hypothesis testing – Which χ^2 test?

χ^2 - goodness-of fit

Measured variable

O_1	O_2	O_3	O_4

O_i is the i^{th} observed value

Expected values are known *a priori*

χ^2 - homogeneity

Measured variable

Group variable stratifying sample	O_1	O_2
	O_3	O_4

Expected values are defined based on H_0, most often column and row totals (proportional distribution)

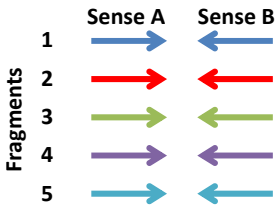
χ^2 - independence

Measured variable 1

Measured variable 2	O_1	O_2
	O_3	O_4

Expected values are defined based on H_0, most often column and row totals (proportional distribution)

Data structure



	A+B	E(x)
1	$O_{1A}+O_{1B}$	?
2	$O_{2A}+O_{2B}$	?
3	$O_{3A}+O_{3B}$	?
4	$O_{4A}+O_{4B}$	?
5	$O_{5A}+O_{5B}$	?

Objectives:

1) Does ligation generate an even distribution of each fragment (A,B,C,D and E).

2) Are sense and anti-sense equally represented for each fragment.

Hypothesis testing – Which χ^2 test?

χ^2 - goodness-of fit

Measured variable

O_1	O_2	O_3	O_4

O_i is the i^{th} observed value

Expected values are known *a priori.* Indeed, they are uniformly distributed (1/5 for each)

χ^2 - homogeneity

Measured variable

Group variable stratifying sample	O_1	O_2
	O_3	O_4

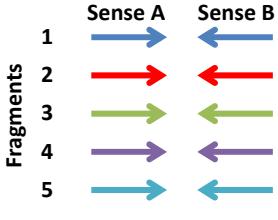
Expected values are defined based on H_0, most often column and row totals (proportional distribution)

χ^2 - independence

Measured variable 1

Measured variable 2	O_1	O_2
	O_3	O_4

Data structure



	A+B	E(x)
1	$O_{1A}+O_{1B}$	1/5
2	$O_{2A}+O_{2B}$	1/5
3	$O_{3A}+O_{3B}$	1/5
4	$O_{4A}+O_{4B}$	1/5
5	$O_{5A}+O_{5B}$	1/5

Objectives: **χ^2 - goodness-of fit**

1) Does ligation generate an even distribution of each fragment (A,B,C,D and E).

2) Are sense and anti-sense equally represented for each fragment.

Hypothesis testing – Which χ^2 test?

χ^2 - goodness-of fit

Measured variable

O_1	O_2	O_3	O_4

O_i is the i^{th} observed value

Expected values are known *a priori*

χ^2 - homogeneity

Measured variable

Group variable stratifying sample	O_1	O_2
	O_3	O_4

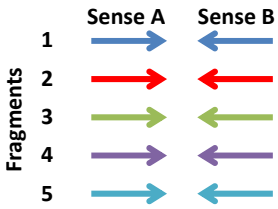
Expected values are defined based on H_0, most often column and row totals (proportional distribution)

χ^2 - independence

Measured variable 1

Measured variable 2	O_1	O_2
	O_3	O_4

Data structure



	A	B
1	O_{1A}	O_{1B}
2	O_{2A}	O_{2B}
3	O_{3A}	O_{3B}
4	O_{4A}	O_{4B}
5	O_{5A}	O_{5B}

Objectives:

1) Does ligation generate an even distribution of each fragment (A,B,C,D and E).

2) Are sense and anti-sense equally represented for each fragment.

Hypothesis testing – Which χ^2 test?

χ^2 - goodness-of fit

Measured variable

O_1	O_2	O_3	O_4

O_i is the i^{th} observed value

Expected values are known *a priori*

χ^2 - homogeneity

Measured variable

Group variable stratifying sample	O_1	O_2
	O_3	O_4

Expected values are defined based on H_0, most often column and row totals (proportional distribution)

χ^2 - independence

Measured variable 1

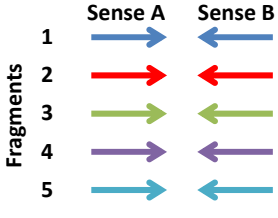
Measured variable 2	O_1	O_2
	O_3	O_4

$$E_{ij} = \frac{n_i \cdot n_j}{n}$$

e.g.

$$E_{1A} = \frac{n_1 \cdot n_A}{n}$$

Data structure



	A	B	Total
1	O_{1A}	O_{1B}	n_1
2	O_{2A}	O_{2B}	n_2
3	O_{3A}	O_{3B}	n_3
4	O_{4A}	O_{4B}	n_4
5	O_{5A}	O_{5B}	n_5
Total	n_A	n_B	n

Objectives:

1) Does ligation generate an even distribution of each fragment (A,B,C,D and E).

2) Are sense and anti-sense equally represented for each fragment.

χ^2 - independence

Hypothesis testing – Which χ^2 test?

χ^2 - goodness-of fit

Measured variable

O_1	O_2	O_3	O_4

O_i is the i^{th} observed value

Expected values are known *a priori*

χ^2 - homogeneity

	Measured variable	
Group variable stratifying sample	O_1	O_2
	O_3	O_4

Expected values are defined based on H_0, most often column and row totals (proportional distribution)

χ^2 - independence

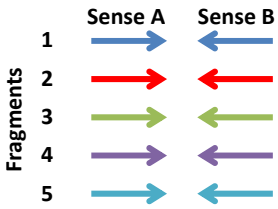
	Measured variable 1	
Measured variable 2	O_1	O_2
	O_3	O_4

$E_{ij} = n_i \cdot 0.5$

e.g.

$E_{1A} = E_{1B} = n_1 \cdot 0.5$

Data structure



	A	B	Total
1	O_{1A}	O_{1B}	n_1
2	O_{2A}	O_{2B}	n_2
3	O_{3A}	O_{3B}	n_3
4	O_{4A}	O_{4B}	n_4
5	O_{5A}	O_{5B}	n_5
Total	n_A	n_B	n

Objectives:

1) Does ligation generate an even distribution of each fragment (A,B,C,D and E).

2) Are sense and anti-sense equally represented for each fragment.

χ^2 - independence

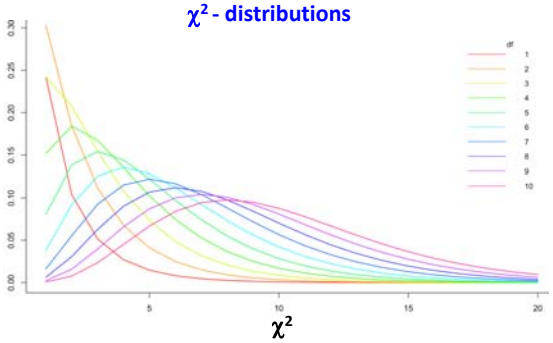
Hypothesis testing - χ^2 distributions

- H_0: Fragments are distributed uniformly (O1) or proportionally (O2)
- H_1: Fragments are not distributed uniformly or proportionally

5. Select the appropriate statistical test
   - Chi-square goodness-of-fit test (objective 1)
   - Chi-square test for independence (objective 2)

} χ^2 is calculated the same way for both tests.

χ^2 - distributions



χ^2

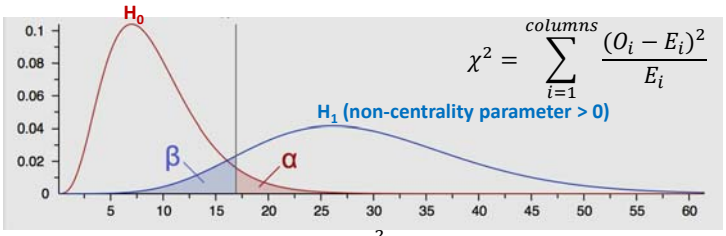
Notice that df is based on the number of categories associated with measured variable. NOT the total number of observations (Contrary to a t-test).

$$\chi^2 = \sum_{i=1}^{columns} \frac{(O_i - E_i)^2}{E_i}, df = columns - 1$$

Hypothesis testing – Power analysis

- H_0: Fragments are distributed uniformly (O1) or proportionally (O2)
- H_1: Fragments are not distributed uniformly or proportionally

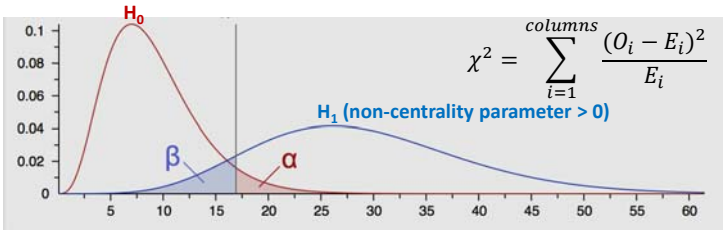
5. Select the appropriate statistical test
   - Chi-square goodness-of-fit test
6. Decide if the Hypothesis testing will be left-tailed, right-tailed, or two tailed test.
   - By definition right-tailed for goodness-of-fit. (χ^2 increases for each error)



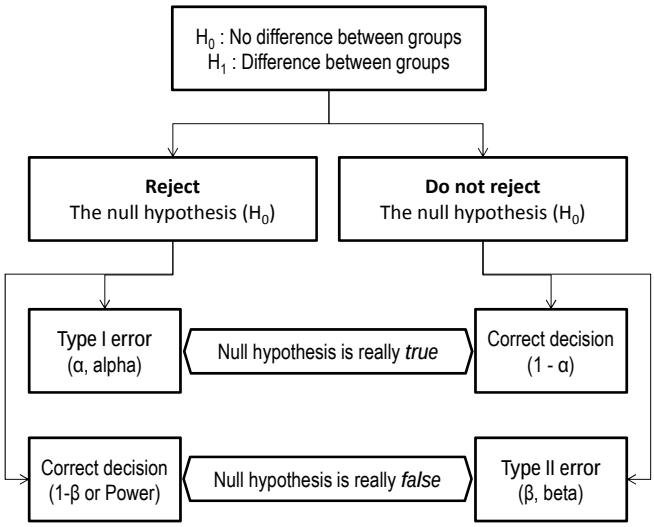
Hypothesis testing – Power analysis

- H_0: Fragments are distributed uniformly (O1) or proportionally (O2)
- H_1: Fragments are not distributed uniformly or proportionally

5. Select the appropriate statistical test
   - Chi-square goodness-of-fit test
6. Decide if the Hypothesis testing will be left-tailed, right-tailed, or two tailed test.
   - By definition right-tailed for goodness-of-fit. (χ^2 increases for each error)
7. State the alpha-risk (α) level : **5% (0.05) (Type-1 error)**
8. State the beta-risk (1-β) level : **90% (0.90) (Type-2 error)**



Type I error, Type II error and Power

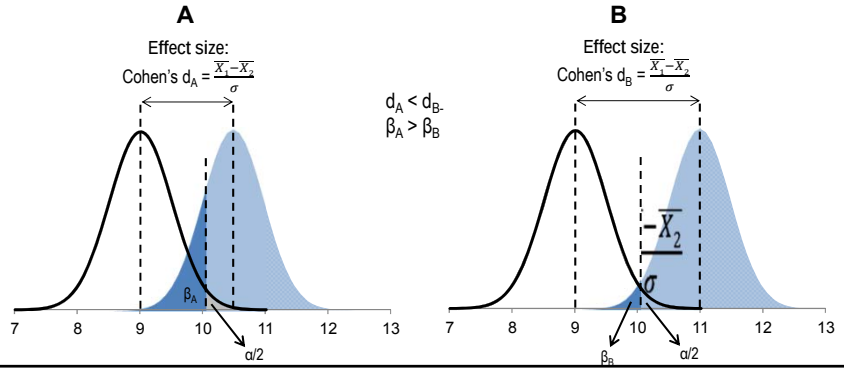


Hypothesis testing – Power analysis

- H_0: Fragments are distributed uniformly (O1) or proportionally (O2)
- H_1: Fragments are not distributed uniformly or proportionally

7. State the alpha-risk (α) level : **5% (0.05) (Type-1 error)**
8. State the beta-risk (1-β) level : **90% (0.90) (Type-2 error)**
9. State or Establish (require prior knowledge) the Effect Size
   - Cohen's interpretation of effect size:

   Low: 0.2 Medium: 0.5 High: 0.8



Hypothesis testing – Power analysis

- H_0: Fragments are distributed uniformly (O1) or proportionally (O2)
- H_1: Fragments are not distributed uniformly or proportionally

7. State the alpha-risk (α) level : **5% (0.05) (Type-1 error)**
8. State the beta-risk (1-β) level : **90% (0.90) (Type-2 error)**
9. State or Establish (require prior knowledge) the Effect Size
   - Interpretation of effect size (depends on df):

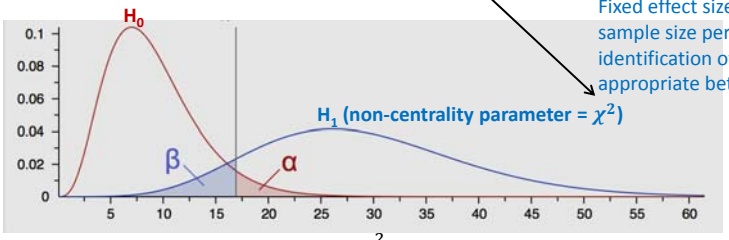
   Low: 0.1 Medium: 0.3 High: 0.5

- To calculate the effect size estimated by the contingency coefficient (ϕ_C):

ϕ_C is less stringent than ϕ because prior knowledge about H_0 distribution improves the strength of the test. Of note, ϕ_C is replaced with ϕ in G*Power or XLSTAT

$$\phi_C = \sqrt{\frac{\chi^2}{\chi^2+N}} \leftrightarrow \chi^2 = \frac{\phi_C^{\,2}\times N}{1-\phi_C^{\,2}} \text{ (identify beta-risk)}$$

Fixed effect size and sample size permit identification of χ^2 with appropriate beta-risk.



Hypothesis testing – Power analysis

- H_0: Fragments are distributed uniformly (O1) or proportionally (O2)
- H_1: Fragments are not distributed uniformly or proportionally

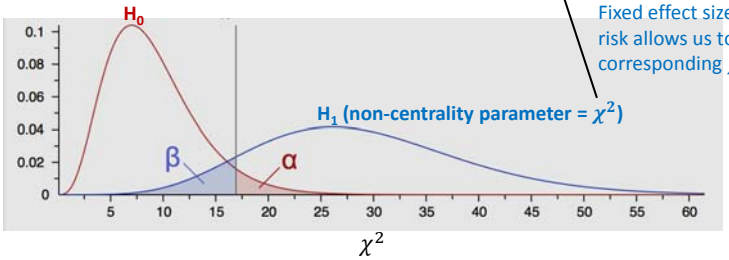
7. State the alpha-risk (α) level : **5% (0.05) (Type-1 error)**
8. State the beta-risk (1-β) level : **90% (0.90) (Type-2 error)**
9. State or Establish (require prior knowledge) the Effect Size
   - Interpretation of effect size (depends on df):

   Low: 0.1 Medium: 0.3 High: 0.5

- To calculate the effect size estimated by the contingency coefficient (ϕ_C):

$$\phi_C = \sqrt{\frac{\chi^2}{\chi^2+N}} \leftrightarrow N = \sqrt{\frac{(1-\phi_C^{\,2})\times\chi^2}{\phi_C^{\,2}}} \text{ (identify sample size)}$$

Fixed effect size and beta-risk allows us to identify corresponding χ^2



Hypothesis testing – Power analysis

- H_0: Fragments are distributed uniformly (O1) or proportionally (O2)
- H_1: Fragments are not distributed uniformly or proportionally

7. State the alpha-risk (α) level : **5% (0.05) (Type-1 error)**
8. State the beta-risk (1-β) level : **90% (0.90) (Type-2 error)**
9. State or Establish (require prior knowledge) the Effect Size
   - Interpretation of effect size (depends on df):

   Low: 0.1 Medium: 0.3 High: 0.5

- To calculate the effect size estimated by the contingency coefficient (ϕ_C):

$$\phi_C = \sqrt{\frac{\chi^2}{\chi^2+N}} \leftrightarrow N = \sqrt{\frac{(1-\phi_c{}^2)\times\chi^2}{\phi_c{}^2}} \text{ (identify sample size)}$$

- Standardize to ϕ_{Max} to avoid effect from altering df.

$$\phi_{Max} = \sqrt{\frac{r-1}{r}}, \text{ (r=rows, r x 1 contingency table)}$$

or

$$\phi_{Max} = \sqrt[4]{\frac{r-1}{r}\times\frac{c-1}{c}} \text{ (r=rows, c=columns, r x c contingency table)}$$

$$\phi_{Standardized} = \frac{\phi_C}{\phi_{Max}}, \text{ ([0,1]=[Independence, Dependence] – interpret like r correlation coefficient)}$$

Hypothesis testing – Power analysis

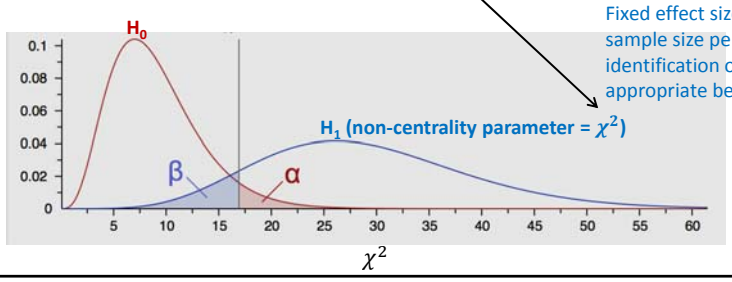
- H_0: Fragments are distributed uniformly (O1) or proportionally (O2)
- H_1: Fragments are not distributed uniformly or proportionally

7. State the alpha-risk (α) level : **5% (0.05) (Type-1 error)**
8. State the beta-risk (1-β) level : **90% (0.90) (Type-2 error)**
9. State or Establish (require prior knowledge) the Effect Size
   - Interpretation of effect size (depends on df):

   Low: 0.1 Medium: 0.3 **High: 0.5**

- To calculate the effect size estimated by phi(ϕ):

$$\phi = \sqrt{\frac{\chi^2}{N}} \leftrightarrow \chi^2 = \phi^2 \times N \text{ (identify beta-risk)}$$



Fixed effect size and sample size permit identification of χ^2 with appropriate beta-risk.

Hypothesis testing – Power analysis

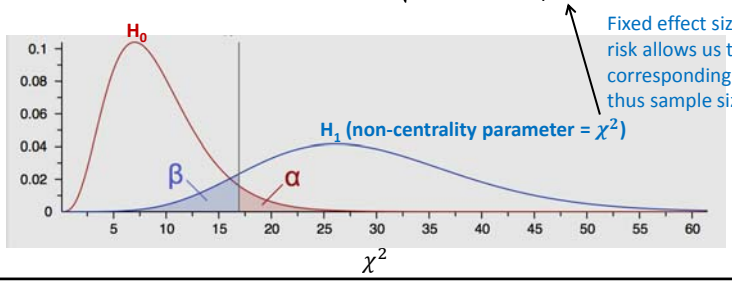
- H_0: Fragments are distributed uniformly (O1) or proportionally (O2)
- H_1: Fragments are not distributed uniformly or proportionally

7. State the alpha-risk (α) level : **5% (0.05) (Type-1 error)**
8. State the beta-risk (1-β) level : **90% (0.90) (Type-2 error)**
9. State or Establish (require prior knowledge) the Effect Size
   - Interpretation of effect size (depends on df):

   Low: 0.1 Medium: 0.3 **High: 0.5**

- To calculate the effect size estimated by phi (ϕ):

$$\phi = \sqrt{\frac{\chi^2}{N}} \leftrightarrow N = \frac{\chi^2}{\phi^2} \text{ (identify sample size)}$$

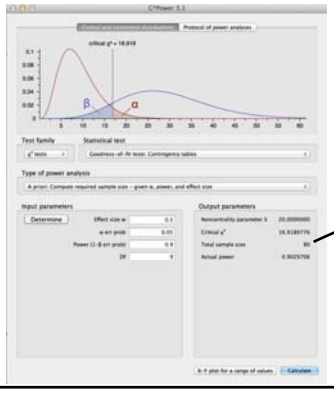


Fixed effect size and beta-risk allows us to identify corresponding χ^2 and thus sample size.

Hypothesis testing – Power analysis

- H_0: Fragments are distributed uniformly (O1) or proportionally (O2)
- H_1: Fragments are not distributed uniformly or proportionally

7. State the alpha-risk (α) level : **5% (0.05) (Type-1 error)**
8. State the beta-risk (1-β) level : **90% (0.90) (Type-2 error)**
9. State or Establish (require prior knowledge) the Effect Size : **ϕ = 0.5**
10. Create Sampling Plan, determine sample size
    - Use G*Power and/or Real Statistics (free software)



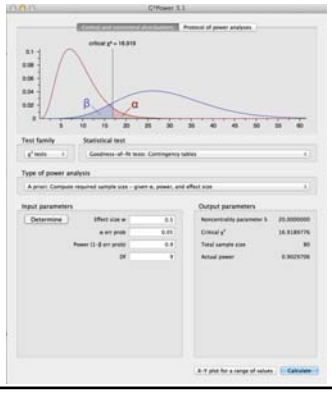
Sample size:

ϕ	N
0.5	80

Hypothesis testing – Power analysis

- H_0: Fragments are distributed uniformly (O1) or proportionally (O2)
- H_1: Fragments are not distributed uniformly or proportionally

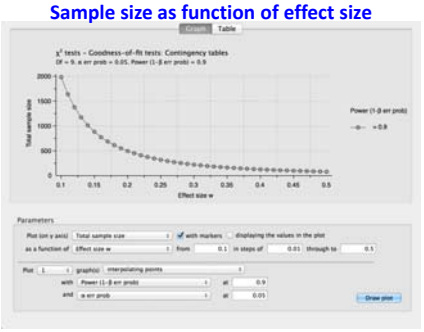
7. State the alpha-risk (α) level : **5% (0.05) (Type-1 error)**
8. State the beta-risk (1-β) level : **90% (0.90) (Type-2 error)**
9. State or Establish (require prior knowledge) the Effect Size : **$\phi = 0.5$**
10. Create Sampling Plan, determine sample size
    - Use G*Power and/or Real Statistics (free software)



Sample size:

ϕ	N
0.1	1983
0.3	221
0.5	80
0.7	41

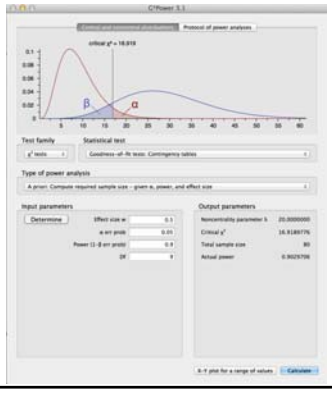
Sample size as function of effect size



Hypothesis testing – Power analysis

- H_0: Fragments are distributed uniformly (O1) or proportionally (O2)
- H_1: Fragments are not distributed uniformly or proportionally

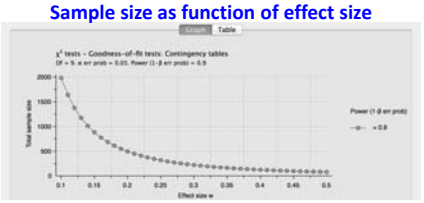
7. State the alpha-risk (α) level : **5% (0.05) (Type-1 error)**
8. State the beta-risk (1-β) level : **90% (0.90) (Type-2 error)**
9. State or Establish (require prior knowledge) the Effect Size : **$\phi = 0.5$**
10. Create Sampling Plan, determine sample size
    - Use G*Power and/or Real Statistics (free software)



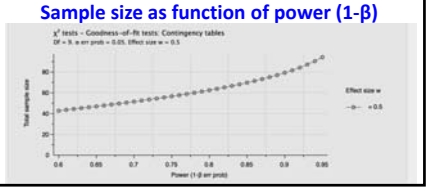
Sample size:

ϕ	N
0.1	1983
0.3	221
0.5	80
0.7	41

Sample size as function of effect size



Sample size as function of power (1-β)



Hypothesis testing – Cohort and data collection

- H_0: Fragments are distributed uniformly (O1) or proportionally (O2)
- H_1: Fragments are not distributed uniformly or proportionally

7. State the alpha-risk (α) level : **5% (0.05) (Type-1 error)**
8. State the beta-risk (1-β) level : **90% (0.90) (Type-2 error)**
9. State or Establish (require prior knowledge) the Effect Size : **ϕ = 0.5**
10. Create Sampling Plan, determine sample size : **N = 80**
11. Gather samples
12. Collect and pre-analyse data

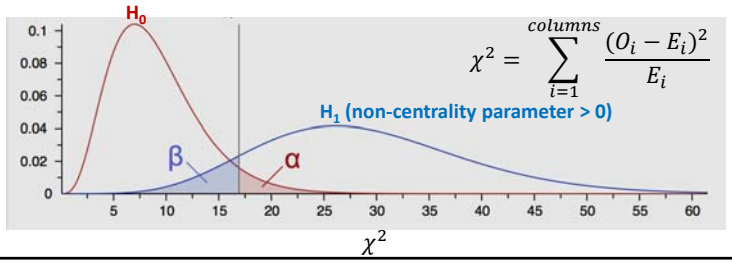
Hypothesis testing – χ^2 test objective 1

- H_0: Fragments are distributed uniformly (O1) or proportionally (O2)
- H_1: Fragments are not distributed uniformly or proportionally

7. State the alpha-risk (α) level : **5% (0.05) (Type-1 error)**
8. State the beta-risk (1-β) level : **90% (0.90) (Type-2 error)**
9. State or Establish (require prior knowledge) the Effect Size : **ϕ = 0.5**
10. Create Sampling Plan, determine sample size : **N = 80**
11. Gather samples
12. Collect and pre-analyse data
13. Calculate the test statistic
    - Obj. 1: Does ligation generate an even distribution of each fragment.

Data from 2016

	A	B
1	3	0
2	12	6
3	6	9
4	15	0
5	0	2



Hypothesis testing - χ^2 test objective 1

- H_0: Fragments are distributed uniformly (O1) or proportionally (O2)
- H_1: Fragments are not distributed uniformly or proportionally

7. State the alpha-risk (α) level : **5% (0.05) (Type-1 error)**
8. State the beta-risk (1-β) level : **90% (0.90) (Type-2 error)**
9. State or Establish (require prior knowledge) the Effect Size : **ϕ = 0.5**
10. Create Sampling Plan, determine sample size : **N = 80**
11. Gather samples
12. Collect and pre-analyse data
13. Calculate the test statistic
    - Obj. 1: Does ligation generate an even distribution of each fragment.

Data from 2016

	A	B
1	3	0
2	12	6
3	6	9
4	15	0
5	0	2

Row Sum →

Observed

	O
1	3
2	18
3	15
4	15
5	2

N=53

Expected

	E
1	53/5
2	53/5
3	53/5
4	53/5
5	53/5

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

$$\frac{(3-10.6)^2}{10.6} + \frac{(18-10.6)^2}{10.6} + \frac{(15-10.6)^2}{10.6} + \frac{(15-10.6)^2}{10.6} + \frac{(2-10.6)^2}{10.6} = 21.25$$

Degrees of freedom (df) = 5-1 = 4

Hypothesis testing - χ^2 test objective 1

- H_0: Fragments are distributed uniformly (O1) or proportionally (O2)
- H_1: Fragments are not distributed uniformly or proportionally

7. State the alpha-risk (α) level : **5% (0.05) (Type-1 error)**
8. State the beta-risk (1-β) level : **90% (0.90) (Type-2 error)**
9. State or Establish (require prior knowledge) the Effect Size : **ϕ = 0.5**
10. Create Sampling Plan, determine sample size : **N = 80**
11. Gather samples
12. Collect and pre-analyse data
13. Calculate the test statistic
    - Obj. 1: Does ligation generate an even distribution of each fragment.

	alpha	
df	0.05	0.01
1	3.84	6.64
2	5.99	9.21
3	7.82	11.34
4	9.49	13.28
5	11.07	15.09
6	12.59	16.81
7	14.07	18.48
8	15.51	20.09
9	16.92	21.67
10	18.31	23.21

Observed

	O
1	3
2	18
3	15
4	15
5	2

N=53

Expected

	E
1	53/5
2	53/5
3	53/5
4	53/5
5	53/5

$\chi^2 = 21.25$

Degrees of freedom (df) = 5-1 = 4

$\chi^2_{Critical} = 9.49 < 21.25$

Statistical Conclusion:
Reject H_0 hypothesis. Samples do not derive from a uniform distribution.

Biological Conclusion:
λ Phage DNA fragments are not cloned into pUC19 in equal proportions.

Hypothesis testing – objective 1 *post hoc*

- **Add-on question:** The fragments are not cloned at equal proportions, but which fragments are significantly different from the uniform distribution?

Hypothesis testing – objective 1 *post hoc*

- **Add-on question:** The fragments are not cloned at equal proportions, but which fragments are significantly different from the uniform distribution?

- This question can be answered with *post-hoc* (Latin, meaning "after this") methods.
- Many *post-hoc* methods are available. The most simple is the repetitive analysis of all possible combinations corrected for Type-I error using **Bonferroni** correction (dividing alpha with the number of combinations tested).
- For objective 1 we could compare each of the fragments with all the other fragments combined (using multiple Chi-square tests).
- An alternative is to consider the observations as a normal distributed random variable and transform them to adjusted residuals following the standardized normal distribution (N(0,1)).
- The latter approach will be scrutinized here.

χ^2-squared - *post hoc* analysis

- Suppose, X is a standard normal random variable (mean = 0 and variance = 1). X ~ N(0,1).
- A sample drawn randomly from X is normally distributed.
- The residuals of the sample will equally be normally distributed this time with a mean of 0. Of note, the variance will differ for each O_{ij}.
  - The variance is inversely correlated with p_i and p_j. The more frequent an observation is for a given measured variable the less variance. We can therefore correct by dividing the residual with $\sqrt{(1-p_i)}$. For frequent observations the adjusted residual will increase more than for infrequent observations.
- The adjusted residuals can be compared to a critical Z-score
  - in excel for two-tailed analysis = NORM.INV(α /2, 0, 1) = NORM.INV(0.025, 0, 1) = 1.96
  - Make correction for multiple comparisons. Bonferroni: $\alpha_{corr} = \alpha/n_{comparisons}$

$E_i = p_i \cdot n$ (Expected) and $p_i = \frac{1}{5}$

Residual: $r_i = \frac{O_i - E_i}{\sqrt{E_i}}$

Adjusted residual $= \frac{O_i - E_i}{\sqrt{E_i(1-p_i)}}$

Hypothesis testing – objective 1 *post hoc*

- **Add-on question:** The fragments are not cloned at equal proportions, but which fragments are significantly different from the uniform distribution?

$E_i = p_i \cdot n$ (Expected) and $p_i = \frac{1}{5}$

Residual: $r_i = \frac{O_i - E_i}{\sqrt{E_i}}$

Adjusted residual $= \frac{O_i - E_i}{\sqrt{E_i(1-p_i)}}$

Conclusion:

Observed

	O_i
1	3
2	18
3	15
4	15
5	2

n=53

Adjusted residuals

	r_{adj}	Size (bp)
1	-2.61	16841
2	2.54	5626
3	1.51	6527
4	1.51	7234
5	-2.95	1275

α = 0.05
Tests = 5
$\alpha_{bonferroni}$ = 0.05/5=0.01
N(0.01/2,0,1)=2.58

Hypothesis testing – objective 1 *post hoc*

- **Add-on question:** The fragments are not cloned at equal proportions, but which fragments are significantly different from the uniform distribution?

$E_i = p_i \cdot n$ (Expected) and $p_i = \frac{1}{5}$

Residual: $r_i = \frac{O_i - E_i}{\sqrt{E_i}}$

Adjusted residual $= \frac{O_i - E_i}{\sqrt{E_i(1-p_i)}}$

Observed

	O_i
1	3
2	18
3	15
4	15
5	2

n=53

Adjusted residuals

	r_{adj}	Size (bp)
1	-2.61	16841
2	2.54	5626
3	1.51	6527
4	1.51	7234
5	-2.95	1275

α = 0.05
Tests = 5
$\alpha_{bonferroni}$ = 0.05/5=0.01
N(0.01/2,0,1)=2.58

Conclusion:
Fragment 2 show a tendency of being over-represented, whereas fragment 1 and 5 are significantly underrepresented.

Why? Find potential causes for your observations
1) Large fragment: ?
2) Small fragment: ?

Hypothesis testing – objective 1 *post hoc*

- **Add-on question:** The fragments are not cloned at equal proportions, but which fragments are significantly different from the uniform distribution?

$E_i = p_i \cdot n$ (Expected) and $p_i = \frac{1}{5}$

Residual: $r_i = \frac{O_i - E_i}{\sqrt{E_i}}$

Adjusted residual $= \frac{O_i - E_i}{\sqrt{E_i(1-p_i)}}$

Observed

	O_i
1	3
2	18
3	15
4	15
5	2

n=53

Adjusted residuals

	r_{adj}	Size (bp)
1	-2.61	16841
2	2.54	5626
3	1.51	6527
4	1.51	7234
5	-2.95	1275

α = 0.05
Tests = 5
$\alpha_{bonferroni}$ = 0.05/5=0.01
N(0.01/2,0,1)=2.58

Conclusion:
Fragment 2 show a tendency of being over-represented, whereas fragment 1 and 5 are significantly underrepresented.

Why? Find potential causes for your observations
1) Large fragment: Transfection efficiency
2) Small fragment: DNA purification

Hypothesis testing - objective 1

- **Apply to your own samples!**

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

$E_i = p_i \cdot n$ (Expected) and $p_i = \frac{1}{5}$

Residual: $r_i = \frac{O_i - E_i}{\sqrt{E_i}}$

Adjusted residual $= \frac{O_i - E_i}{\sqrt{E_i(1-p_i)}}$

Observed

	O
1	O_1
2	O_2
3	O_3
4	O_4
5	O_5

Expected

	E
1	E_1
2	E_2
3	E_3
4	E_4
5	E_5

Hypothesis testing - χ^2 test objective 2

- H_0: Fragments are distributed uniformly (O1) or proportionally (O2)
- H_1: Fragments are not distributed uniformly or proportionally

7. State the alpha-risk (α) level : **5% (0.05) (Type-1 error)**
8. State the beta-risk (1-β) level : **90% (0.90) (Type-2 error)**
9. State or Establish (require prior knowledge) the Effect Size : **ϕ = 0.5**
10. Create Sampling Plan, determine sample size : **N = 80**
11. Gather samples
12. Collect and pre-analyse data
13. Calculate the test statistic
    - Obj. 2: Are sense and anti-sense equally represented for each fragment.

Observations

	A	B	Total
1	3	0	3
2	12	6	18
3	6	9	15
4	15	0	15
5	0	2	2
Total	36	17	53

50:50?

Expected

	A	B
1	E_{1A}	E_{1B}
2	E_{2A}	E_{2B}
3	E_{3A}	E_{3B}
4	E_{4A}	E_{4B}
5	E_{5A}	E_{5B}

50:50

$$E_{ij} = n_i \cdot 0.5$$

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

Hypothesis testing - χ^2 test objective 2

- H_0: Fragments are distributed uniformly (O1) or proportionally (O2)
- H_1: Fragments are not distributed uniformly or proportionally

7. State the alpha-risk (α) level : **5% (0.05) (Type-1 error)**
8. State the beta-risk (1-β) level : **90% (0.90) (Type-2 error)**
9. State or Establish (require prior knowledge) the Effect Size : **ϕ = 0.5**
10. Create Sampling Plan, determine sample size : **N = 80**
11. Gather samples
12. Collect and pre-analyse data
13. Calculate the test statistic
    - Obj. 2: Are sense and anti-sense equally represented for each fragment.

Observations

	A	B	Total
1	3	0	3
2	12	6	18
3	6	9	15
4	15	0	15
5	0	2	2
Total	36	17	53

50:50?

Expected

	A	B
1	1.5	1.5
2	9	9
3	7.5	7.5
4	7.5	7.5
5	1	1

50:50
40% E_{ij}<5

$$E_{ij} = n_i \cdot 0.5$$

$$\chi^2 = \sum_{j=1}^{rows} \sum_{i=1}^{columns} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 22.6$$

Hypothesis testing - χ^2 test objective 2

- H_0: Fragments are distributed uniformly (O1) or proportionally (O2)
- H_1: Fragments are not distributed uniformly or proportionally

7. State the alpha-risk (α) level : **5% (0.05) (Type-1 error)**
8. State the beta-risk (1-β) level : **90% (0.90) (Type-2 error)**
9. State or Establish (require prior knowledge) the Effect Size : **ϕ = 0.5**
10. Create Sampling Plan, determine sample size : **N = 80**
11. Gather samples
12. Collect and pre-analyse data
13. Calculate the test statistic
    - Obj. 2: Are sense and anti-sense equally represented for each fragment.

Observations

	A	B	Total
1+5	3	2	5
2	12	6	18
3	6	9	15
4	15	0	15
Total	36	17	53

50:50?

Expected

	A	B
1+5	2.5	2.5
2	9	9
3	7.5	7.5
4	7.5	7.5

50:50
25% E_{ij}<5

$$E_{ij} = n_i \cdot 0.5$$

$$\chi^2 = \sum_{j=1}^{rows} \sum_{i=1}^{columns} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 17.8$$

Hypothesis testing - χ^2 test objective 2

- H_0: Fragments are distributed uniformly (O1) or proportionally (O2)
- H_1: Fragments are not distributed uniformly or proportionally

7. State the alpha-risk (α) level : **5% (0.05) (Type-1 error)**
8. State the beta-risk (1-β) level : **90% (0.90) (Type-2 error)**
9. State or Establish (require prior knowledge) the Effect Size : **ϕ = 0.5**
10. Create Sampling Plan, determine sample size : **N = 80**
11. Gather samples
12. Collect and pre-analyse data
13. Calculate the test statistic
    - Obj. 2: Are sense and anti-sense equally represented for each fragment.

Observations

	A	B	Total
1+5	3	2	5
2	12	6	18
3	6	9	15
4	15	0	15
Total	36	17	53

50:50?

Expected

	A	B
1+5	2.5	2.5
2	9	9
3	7.5	7.5
4	7.5	7.5

50:50
25% E_{ij}<5

$$E_{ij} = n_i \cdot 0.5$$

$$\chi^2 = \sum_{j=1}^{rows} \sum_{i=1}^{columns} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 17.8$$

Degré de libérté (df) = (ligne-1)*(colonne-1) = (2-1)*(4-1) = 3

$\chi^2_{Critical} = 7.82 < 17.8$ => Reject H_0

χ^2-squared - objective 2 *post hoc*

- Suppose, X is a standard normal random variable (mean = 0 and variance = 1). X ~ N(0,1).
- A sample drawn randomly from X is normally distributed.
- The residuals of the sample will equally be normally distributed this time with a mean of 0. Of note, the variance will differ for each O_{ij}.
  - The variance is inversely correlated with p_i and p_j. The more frequent an observation is for a given measured variable the less variance. We can therefore correct by dividing the residual with $\sqrt{(1-p_i)(1-p_j)}$. For frequent observations the adjusted residual will increase more than for infrequent observations.
- The adjusted residuals can be compared to the critical Z-score
  - in excel for two-tailed analysis = NORM.INV(α /2, 0, 1) = NORM.INV(0.025, 0, 1) = 1.96
  - Make correction for multiple comparisons. Bonferroni: $\alpha_{corr} = \alpha/n_{comparisons}$

$E_{ij} = n_i \cdot pj$ (Expected) and $p_i = \frac{n_i}{n}$, p_j=0.5

Residual: $r_{ij} = \frac{O_{ij}-E_{ij}}{\sqrt{E_{ij}}}$

Adjusted residual = $\frac{O_{ij}-E_{ij}}{\sqrt{E_{ij}(1-p_i)(1-p_j)}}$

		M_j		
		j=1	j=2	
M_i	i=1	O_{ij}	O_{ij}	$n_{i=1}$
	i=2	O_{ij}	O_{ij}	$n_{i=2}$
		$n_{j=1}$	$n_{j=2}$	N

M = Measure (variable)
O = Observation
n = Row, column or total sums

Hypothesis testing - objective 2 *post hoc*

- **Add-on question:** Which fragment has a skewed sense/anti-sense distribution?

$E_{ij} = n_i \cdot pj$ (Expected) and $p_i = \frac{n_i}{n}$, p_j=0.5

Residual: $r_{ij} = \frac{O_{ij} - E_{ij}}{\sqrt{E_{ij}}}$

Adjusted residual = $\frac{O_{ij} - Ei_i}{\sqrt{E_{ij}(1-p_i)(1-p_j)}}$

Observed

	A	B	Total
1	3	0	3
2	12	6	18
3	6	9	15
4	15	0	15
5	0	2	2
Total	36	17	53

Adjusted residuals

	A	B
1	1.78	-1.78
2	1.74	-1.74
3	-0.91	0.91
4	4.57	-4.57
5	-1.44	1.44

α = 0.05
Tests = 10
$\alpha_{bonferroni}$ = 0.05/10=0.005
N(0.005/2,0,1)=2.81

Conclusion:

Hypothesis testing - objective 2 *post hoc*

- **Add-on question:** Which fragment has a skewed sense/anti-sense distribution?

$E_{ij} = n_i \cdot pj$ (Expected) and $p_i = \frac{n_i}{n}$, p_j=0.5

Residual: $r_{ij} = \frac{O_{ij} - E_{ij}}{\sqrt{E_{ij}}}$

Adjusted residual = $\frac{O_{ij} - Ei_i}{\sqrt{E_{ij}(1-p_i)(1-p_j)}}$

Observed

	A	B	Total
1	3	0	3
2	12	6	18
3	6	9	15
4	15	0	15
5	0	2	2
Total	36	17	53

Adjusted residuals

	A	B
1	1.78	-1.78
2	1.74	-1.74
3	-0.91	0.91
4	4.57	-4.57
5	-1.44	1.44

α = 0.05
Tests = 10
$\alpha_{bonferroni}$ = 0.05/10=0.005
N(0.005/2,0,1)=2.81

Conclusion:
Fragment 4 is significantly more represented in the sense direction (sense A) compared to the anti-sense.

Why? Find a cause for your finding.
1)

Hypothesis testing - objective 2 *post hoc*

- **Add-on question:** Which fragment has a skewed sense/anti-sense distribution?

$E_{ij} = n_i \cdot pj$ (Expected) and $p_i = \frac{n_i}{n}$, p_j=0.5

Residual: $r_{ij} = \frac{O_{ij} - E_{ij}}{\sqrt{E_{ij}}}$

Adjusted residual = $\frac{O_{ij} - Ei_j}{\sqrt{E_{ij}(1-p_i)(1-p_j)}}$

Observed

	A	B	Total
1	3	0	3
2	12	6	18
3	6	9	15
4	15	0	15
5	0	2	2
Total	36	17	53

Adjusted residuals

	A	B
1	1.78	-1.78
2	1.74	-1.74
3	-0.91	0.91
4	4.57	-4.57
5	-1.44	1.44

α = 0.05
Tests = 10
$\alpha_{bonferroni}$ = 0.05/10=0.005
N(0.005/2,0,1)=2.81

Conclusion:
Fragment 4 is significantly more represented in the sense direction (sense A) compared to the anti-sense.

Why? Find a cause for your finding.
1) Toxicity

Hypothesis testing - objective 2

- **Apply to your own samples!**

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

$E_{ij} = n_i \cdot p_j$ (expected) and $p_i = \frac{n_i}{n}$, p_j=0.5

Residual: $r_{ij} = \frac{O_{ij} - E_{ij}}{\sqrt{E_{ij}}}$

Adjusted residual = $\frac{O_{ij} - E_{ij}}{\sqrt{E_{ij}(1-p_i)(1-p_j)}}$

Observed

	A	B	Total
1	O_{1A}	O_{1B}	n_1
2	O_{2A}	O_{2B}	n_2
3	O_{3A}	O_{3B}	n_3
4	O_{4A}	O_{4B}	n_4
5	O_{5A}	O_{5B}	n_5
Total	n_A	n_B	n

Estimated

	A	B
1	E_{1A}	E_{1B}
2	E_{2A}	E_{2B}
3	E_{3A}	E_{3B}
4	E_{4A}	E_{4B}
5	E_{5A}	E_{5B}