Data Analysis in Astronomy and Physics

Lecture 7: Hypothesis Testing for Categorical Variables

M. Röllig

Decision errors

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	✓	Type 1 error
	H_A true	Type 2 error	✓

- Type 1 error is rejecting H_0 when H_0 is true. (False positive)
- Type 2 error is failing to reject H_0 when H_A is true. (False negative)
- We (almost) never know if H_0 or H_A is true, but we need to consider all possibilities.

Type 1 error rate α

- We reject H_0 when the p-value is less than 0.05 (α=0.05).
- This means that, for those cases where H_0 is actually true, we do not want to incorrectly reject it more than 5% of those times.
- In other words, when using a 5% significance level there is about 5% chance of making a Type 1 error if the null hypothesis is true.

Out[ ]//TraditionalForm=

$$\mathcal{P}(\text{Type 1 error} \mid H_0 \text{ true}) = \alpha$$

- This is why we prefer a small value of α - increasing α increases the Type 1 error rate.

Type 2 error rate β

If the alternative hypothesis is actually true, what is the chance that we make a Type 2 error, i.e. we fail to reject the null hypothesis even when we should reject it?

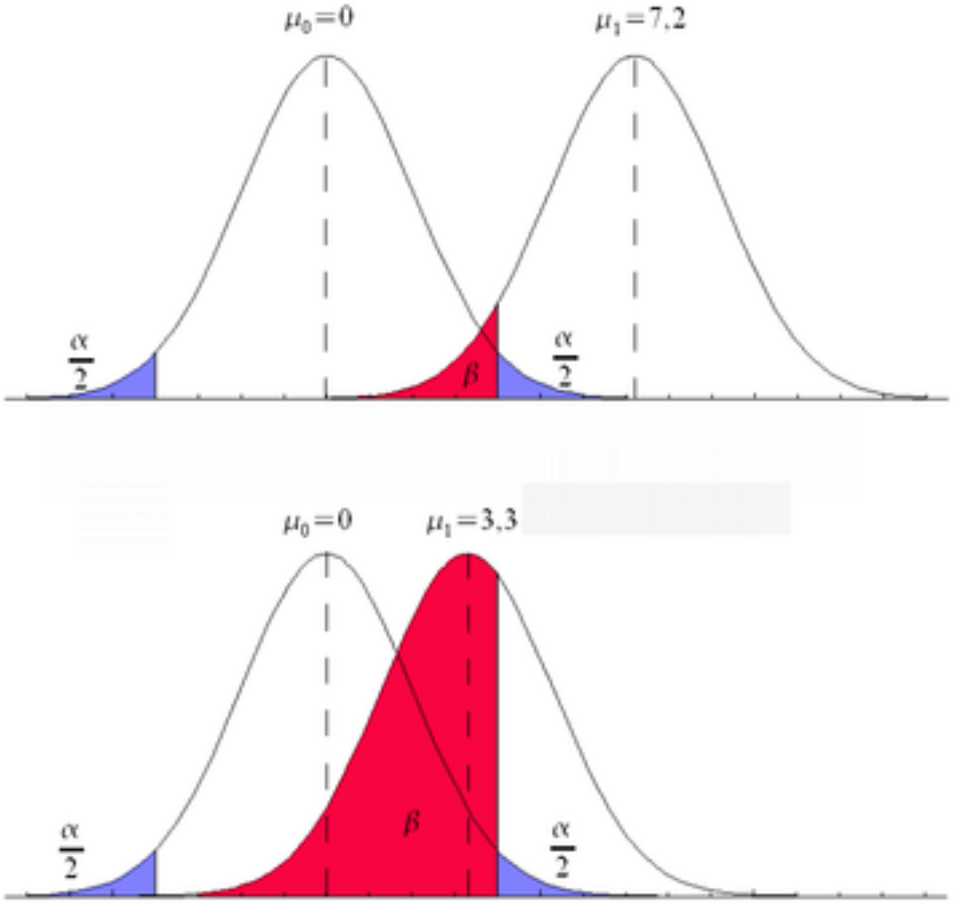
- The answer is not obvious.
- If the true population average is very close to the null value, it will be difficult to detect a difference (and reject H_0).
- If the true population average is very different from the null value, it will be easier to detect a difference.
- Clearly, β depends on the **effect size δ**, difference between point estimate and null value.

Decision error

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	$1-\alpha$	Type 1 error, α
	H_A true	Type 2 error, β	$1-\beta$

- **Type 1 error** is rejecting H_0 when you shouldn't have, and the probability of doing so is $\boldsymbol{\alpha}$ (significance level)
- **Type 2 error** is failing to reject H_0 when you should have, and the probability of doing so is $\boldsymbol{\beta}$.
- **Power** of a test is the probability of correctly rejecting H_0, and the probability of doing so is $\mathbf{1 - \boldsymbol{\beta}}$

Decision error

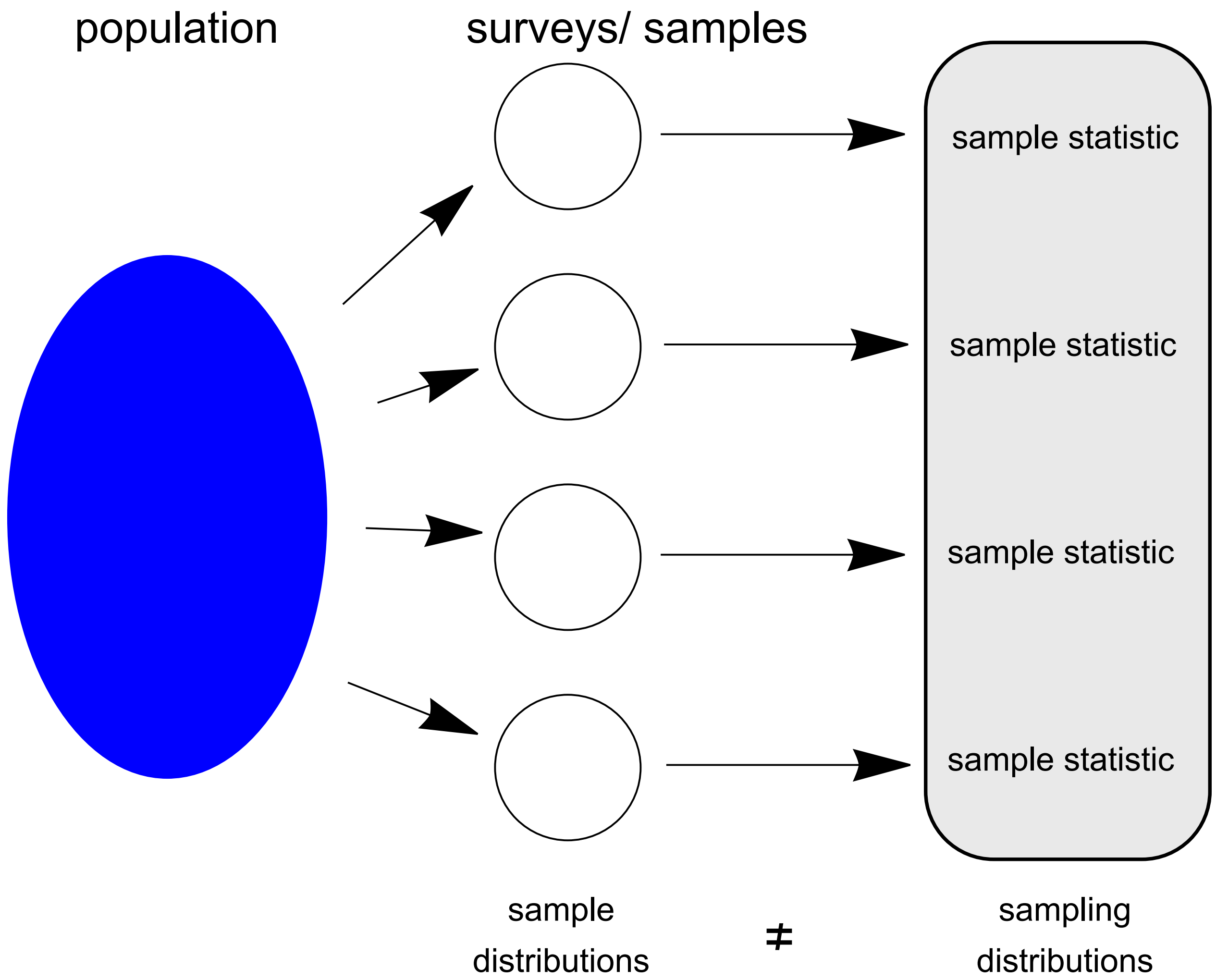


Red: Type 2 error

The type 2 error depends on the value of the true parameter μ. The clo
the value assumed under the null hypothesis the larger β becomes.

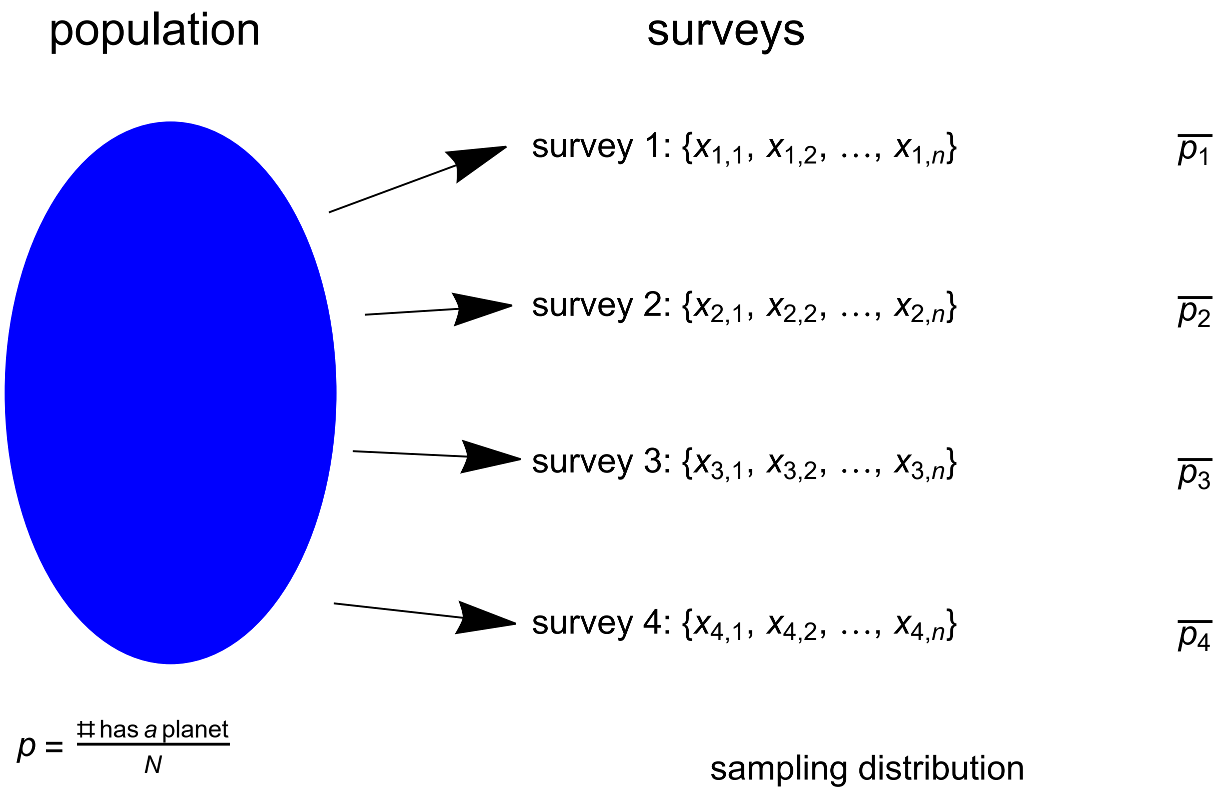
Sampling variability & CLT for proportions

Example: We have access to 4 exoplanet surveys. Each survey lists stellar system together with a yes or no, depending on whether this system hosts at least a planet.


population
surveys/ samples
sample statistic
sample statistic
sample statistic
sample statistic
sample
distributions
≠
sampling
distributions

Sampling variability & CLT for proportions

The $x_{i,j}$ are the categorical variable of the j-th observation in the i-th survey. It can be either yes or no. We summarize the categorical values for each survey to **convert it into the proportion** of observations that yielded the category we are interested in, e.g. % of systems with planets.



sampling distribution

mean($\overline{p}$) ≈ μ

CLT for proportions

The distribution of sample proportions is nearly normal, centered at the population proportion, and with a standard error inversely proportional to the sample size.

Out[∘]//TraditionalForm=

$$\boxed{\overline{p} \approx \mathcal{N}\left(\text{mean} = \text{p}, \text{SE}= \sqrt{\frac{p(1-p)}{n}}\right)}$$

This is related to the Binomial distribution (the category of interest is the “success” event).

- Let $B = B(n, p)$ denote the binomial random variable with n trials and probability p
- The B is the number of successes in a random sample
- Thus the sample proportion $\overline{p}$ = (# in category) / (# in sample) = $\frac{1}{n} B$
- We know that approximately $B(n, p) \approx \mathcal{N}(\text{np}, \sqrt{\text{np}(1-p)})$
- It follows $\overline{p} = \frac{1}{n} B \approx \frac{1}{n} \mathcal{N}(\text{np}, \sqrt{\text{np}(1-p)}) = \mathcal{N}(\frac{\text{np}}{n}, \frac{1}{n} (\sqrt{\text{np}(1-p)})) = \mathcal{N}\left(p, \sqrt{\left(\frac{p(1-p)}{n}\right)}\right)$

CLT for proportions

Example

70% of all stars are classified as multiple systems. If you were to randomly select 200 stars from the sky, what is the probability, that at least 90% of stars in your sample will be part of a multi-star system?

$p = 0.7,\ n = 200,\ \mathcal{P}_r(\overline{p} > 0.9) = ?$

1. conditions met? independent sample and < 10% of all stars $\longrightarrow$ YES
2. $n\,p = 200 * 0.7 = 140$ and $200 * (1 - 0.7) = 60$

$\overline{p} \approx \mathcal{N}\left(\text{mean} = 0.7,\ \text{SE} = \sqrt{\left(\frac{0.7*0.3}{200}\right)} = 0.0324037\right) \quad Z = \frac{0.9-0.7}{0.03} = 6.66667, \quad \mathcal{P}_r(Z > 6.67) \approx 0.$

CI for a proportion

point estimate ± margin of error

Out[ ]//TraditionalForm=

$$\overline{p} \pm z^{*}\, \mathrm{SE}_{\overline{p}}$$

Standard error for a proportion for calculating a CI

Out[ ]//TraditionalForm=

$$\mathrm{SE}_{\overline{p}} = \sqrt{\frac{\overline{p}\,(1 - \overline{p})}{n}}$$

Hypothesis testing for a single proportion

1. Set the hypotheses: $H_0 : p = null\ value$

   $H_A : p <$ or $>$ or $\neq$ $null\ value$

2. Calculate the point estimate: $\overline{p}$
3. Check conditions:

   3.1. **Independence**: Sampled observations must be independent (random sample/assignment & if sampling without replacement, n < 10% of population)

   3.2. **Sample size/skew:** n p≥10, n(1-p)≥10, larger if the population distribution is very skewed.

4. Draw sampling distribution, shade p-value, calculate test statistic $Z = \frac{\overline{p}-p}{\text{SE}}$, SE = $\sqrt{\frac{p(1-p)}{n}}$
5. Make a decision, and interpret it in context of the research question:
   - **if p-value <α, reject H_0;** the data provide convincing evidence for H_A
   - **if p-value >α, fail to reject H_0;** the data **do not** provide convincing evidence for H_A

$\overline{p}$ versus p

sample proportion: $\overline{p}$, population proportion: p

	confidence interval	hypothesis test
success – failure condition	$n\,\overline{p} \geq 10$ $n(1-\overline{p}) \geq 10$	$n\,p \geq 10$ $n\,(1-p) \geq 10$
standard error	$SE = \sqrt{\frac{\overline{p}\,(1-\overline{p})}{n}}$	$SE = \sqrt{\frac{p\,(1-p)}{n}}$

Example

A 2013 Pew Research poll found that 60% of 1,983 randomly sampled American adults believe in evolution. Does this provide convincing evidence that majority of Americans believe in evolution?

$H_0 : p = 0.5 \quad \overline{p} = 0.6$

$H_A : p > 0.5 \quad n = 1983$

1. conditions? independence: 1983<10% of all Americans & random sample
2. sample size/skew: 1983×0.5=991.5>10

$\overline{p} \approx \mathcal{N}\left(\text{mean} = 0.5,\ \text{SE} = \sqrt{\frac{0.5\times 0.5}{1983}} = 0.0112282\right)$

$Z = \frac{0.6-0.5}{0.0112} = 8.92857$ p-value=$\mathcal{P}_r$(Z>8.93)= almost 0 $\longrightarrow$ reject H_0

Estimating the difference between two proportions

parameter of interest: $p_1 - p_2$
point estimate: $\overline{p}_1 - \overline{p}_2$

point estimate ± margin of error

Out[∘]//TraditionalForm=

$$(\overline{p}_1 - \overline{p}_2) \pm z^* \, \mathrm{SE}_{(\overline{p}_1 - \overline{p}_2)}$$

Standard error for difference between two proportions, for calculating a confidence interval:

Out[∘]//TraditionalForm=

$$\mathrm{SE}_{\overline{p}} = \sqrt{\frac{\bar{p}_1 (1 - \bar{p}_1)}{n_1} + \frac{\bar{p}_2 (1 - \bar{p}_2)}{n_2}}$$

Conditions for comparing two independent proportions

- **Independence:**
  - **within groups**: sampled observations must be independent within each group
    - random sample/assignment
    - if sampling without replacement, n < 10% of population)
  - **between groups:** the two groups must be independent of each other (non-paired)
- **Sample size/skew:** Each sample should meet the success-failure condition:
  - $n_1 p_1 \geq 10$ and $n_1(1 - p_1) \geq 10$
  - $n_2 p_2 \geq 10$ and $n_2(1 - p_1) \geq 10$

Hypothesis tests for comparing two proportions

A SurveyUSA poll asked respondents whether any of their children have ever been the victim of bullying. Also recorded on this survey was the gender of the respondent (the parent). Below is the distribution of responses by gender of the respondent.

	Male	Female
Yes	34	61
No	52	61
Not sure	4	0
Total	90	122
$\overline{p}$	0.38	0.5

$H_0 : p_{\text{male}} - p_{\text{female}} = 0$

$H_A : p_{\text{male}} - p_{\text{female}} \neq 0$

$\overline{p}$ versus p

	confidence interval	hypothesis test
success – failure condition	$n\,\overline{p} \geq 10$ $n(1-\overline{p}) \geq 10$	$n\,p \geq 10$ $n\,(1-p) \geq 10$
standard error	$SE = \sqrt{\frac{\overline{p}(1-\overline{p})}{n}}$	$SE = \sqrt{\frac{p\,(1-p)}{n}}$

When you're dealing with a confidence interval, use the observed counts and observed proportions $\overline{p}$. When you're dealing with a hypothesis test, use expected counts and expected proportions p.

two proportions: $\overline{p}$ versus p

	confidence interval	hypothesis test
success – failure condition	$n_1\ \overline{p}_1 \geq 10,\ n_2\ \overline{p}_2 \geq 10$ $n_1(1-\overline{p}_1) \geq 10,\ n_2(1-\overline{p}_2) \geq 10$	$H_0 : p_1 = p_2$
standard error	$SE = \sqrt{\frac{\overline{p}_1(1-\overline{p}_1)}{n_1} + \frac{\overline{p}_2(1-\overline{p}_2)}{n_2}}$	

The null hypothesis states that the two population proportions should be equal to each other. But equal to what value?

Pooled proportions

$H_0: p_1 = p_2 = ?$

Out[∘]//TraditionalForm=

$$\boxed{\text{pooled proportion :} \qquad \overline{p}_{\text{pool}} = \frac{\text{total successes}}{\text{total } n} = \frac{\sharp \text{ of successes}_1 + \sharp \text{ of successes}_2}{n_1 + n_2}}$$

$\overline{p}_{\text{pool}} = \frac{34+61}{90+122} \approx 0.448113$

two proportions: $\overline{p}$ versus p

	observed confidence interval	***expected*** hypothesis test
success – failure condition	$n_1\ \overline{p}_1 \geq 10,\ n_2\ \overline{p}_2 \geq 10$ $n_1(1-\overline{p}_1) \geq 10,\ n_2(1-\overline{p}_2) \geq 10$	$n_1\ \overline{p}_{\text{pool}} \geq 10,\ n_2\ \overline{p}_{\text{pool}} \geq 10$ $n_1(1-\overline{p}_{\text{pool}}) \geq 10,\ n_2(1-\overline{p}_{\text{pool}}) \geq 10$
standard error	$SE = \sqrt{\frac{\overline{p}_1(1-\overline{p}_1)}{n_1} + \frac{\overline{p}_2(1-\overline{p}_2)}{n_2}}$	$SE = \sqrt{\frac{\overline{p}_{\text{pool}}\left(1-\overline{p}_{\text{pool}}\right)}{n_1} + \frac{\overline{p}_{\text{pool}}\left(1-\overline{p}_{\text{pool}}\right)}{n_2}}$

Example

Conduct a hypothesis test, at 5% significance level, evaluating if males and females are equally likely to answer “Yes” to the question about whether any of their children have ever been the victim of bullying.

	Male	Female
Total	90	122
$\overline{p}$	0.38	0.5
$\overline{p}_{pool}$	0.45	

$H_0 : p_{\text{male}} = p_{\text{female}} = 0, \quad H_0 : p_{\text{male}} \neq p_{\text{female}} \neq 0$

$$(\overline{p}_{\text{male}} - \overline{p}_{\text{female}}) \approx \mathcal{N}\left(\text{mean} = 0, \text{SE} = \sqrt{\frac{(0.45\times0.55)}{90} + \frac{(0.45\times0.55)}{122}} \approx 0.0691\right)$$

point estimate $\overline{p}_{\text{male}} - \overline{p}_{\text{female}} = 0.38 - 0.5 = -0.12$

Example

H_0:μ=0

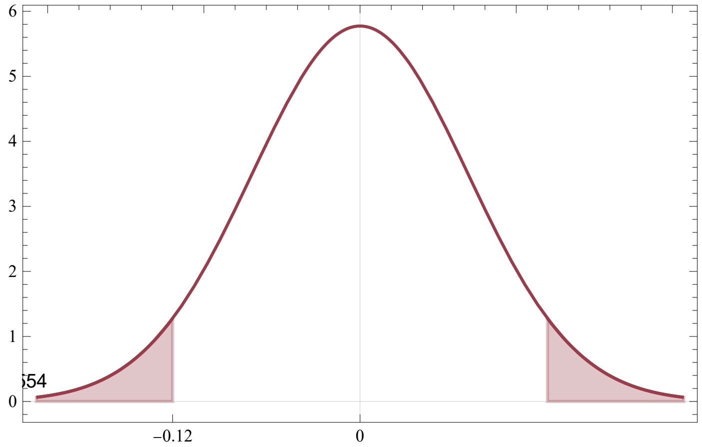
H_A:μ≠0

$\overline{x}$=−0.12

s=0.0691

n=1

$\overline{x}_{\text{diff}} \approx \mathcal{N}(\text{mean}=0;\ \text{SE}=\frac{0.0691}{\sqrt{1}} \approx 0.0691)$



$Z=\frac{-0.12-0}{\frac{0.0691}{\sqrt{1}}}=-1.73661$

p−value=0.0824554

Small sample proportions

The tests on proportions relied on the mean success-failure condition, i.e. the fact that the Binomial distribution approaches the Normal distribution, which we tested against using the Z-score. What if the S-F condition is not met?

We will discuss two possible approaches: **simulation** and **exact probabilities**.

Small sample proportions - Simulation

http://en.wikipedia.org/wiki/Paul_the_Octopus

Paul the Octopus predicted 8 World Cup games, and predicted them all correctly. Does this provide convincing evidence that Paul actually has psychic powers, i.e. that he does better than just randomly guessing?

$H_0 : p = 0.5$, $H_A : p > 0.5$, n=8, $\overline{p} = 1$

1. independence: Yes, we can assume the guesses were independent
2. sample size/skew: $8 \times 0.5 = 4 < 10$ NO, distribution of sample proportions can not be assumed to be nearly normal.

Set up a simulation scheme that assumes the null hypothesis is true!

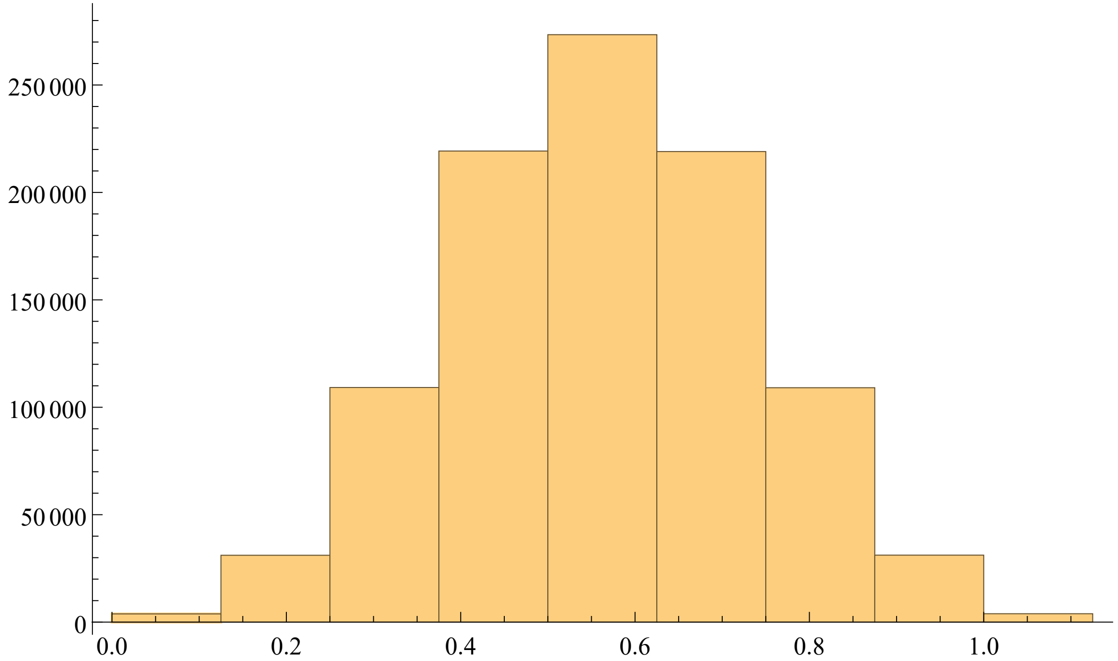
Small sample proportions - Simulation

$H_0 : p = 0.5$, $H_A : p > 0.5$, n=8, $\overline{p} = 1$

E.g. : Create 8 random real numbers between 0 and 1 (0...0.49: failure, 0.5...1.0: success) or flip a coin 8 times and assign head as success.

- one simulation: record the proportion of successes $\longrightarrow \overline{p}_{\text{sim }1}$
- repeat the simulation N times $\longrightarrow \overline{p}_{\text{sim }1}, \ldots, \overline{p}_{\text{sim }N}$
- calculate the percentage of simulations where the simulated proportions of successes is at least as extreme as the observed proportion

Small sample proportions - Simulation



Out[ ]= simulations with $\overline{p}$=1: 3902 of 1 000 000 simulations,   p-value=0.003902

Note, that the bins run from [a, b), excluding the upper limit and including the lower limit.

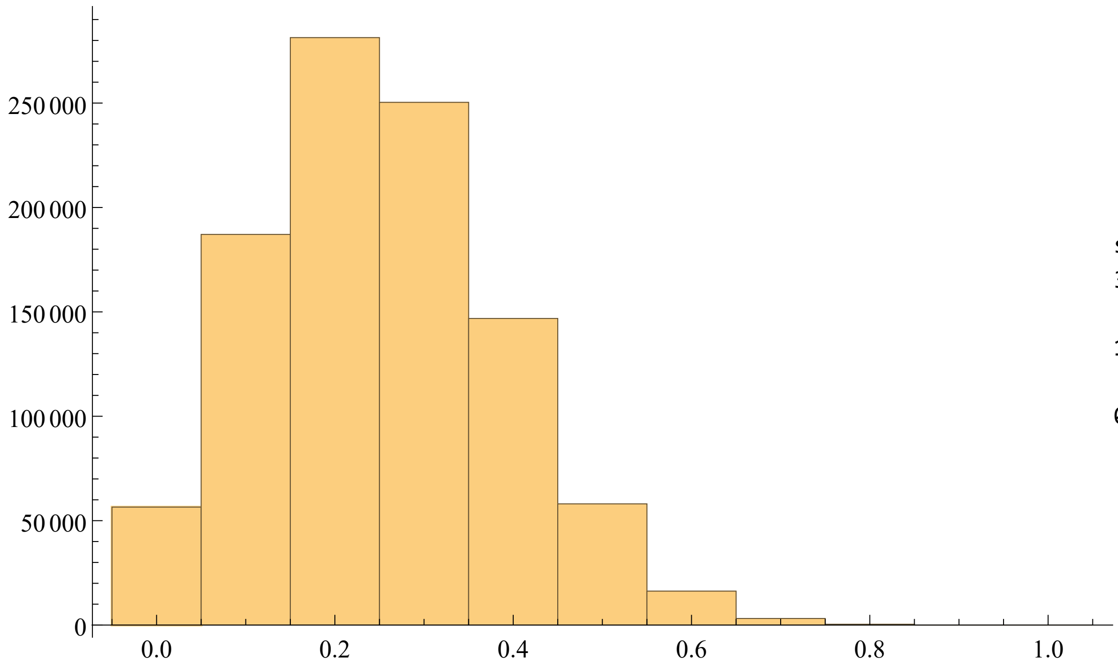
Small sample proportions - Exact

Probability of exactly 8 successes in 8 trials: $\binom{8}{8} 0.5^8 (1-0.5)^{8-8} = 0.00390625$

Small sample proportions - Example

A community based classification of YSO SEDs (Young stellar objects, spectral energy distribution) is tested with a small sample of 10 persons. Each person is shown the SEDs of an Class-I YSO and asked to classify it into one of 4 classes (0,I,II,III) based on the shape of the SED. (See: http://en.wikipedia.org/wiki/Young_stellar_object, and Lada, C. J. (1987), IAU Symposium 115: Star Forming Regions, 115, 1). 7 of the 10 persons chose the correct class. What are the hypotheses for evaluating whether these data provide convincing evidence that people do better than random guessing when it comes to recognizing YSO classes?

Out[◦]=



simulations with $\overline{p}\geq 7/10=0.7$: 3552 of 1 000 000 simulations, p-value= 0.003552

$H_0: p = 1/4 = 0.25,\ \ H_A: p > 1/4 = 0.25$

$$\mathcal{P}r(\overline{p} \geq 7/10 \mid H_0: p = 0.25) = \sum_{i=7}^{10} \binom{10}{i} 0.25^i\,(1-0.25)^{10-i} = 0.00350571$$

The chance to find the date if the persons were randomly guessing is only .35%, so we discard H_0.

Chi-square (χ^2) goodness of fit (GOF) test

Condition: one categorical variables with more than 2 levels.

Example (numbers made up!)

The distribution of spectral types of stars in a globular cluster is:

spectral type	G,K,M	F	O	B	other
% in population	80.29	12.06	0.79	2.92	3.94

Distribution of a sample of 2500 stars of the same globular cluster.

spectral type	G,K,M	F	O	B	other
observed	1920	347	19	84	130

Example (numbers made up!)

You need to decide whether the observed sample is a random sample of the globular cluster or whether it is a biased sample.

H_0 (nothing is going on)	Stars selected are a simple random sample from the population of all GC stars. The observed counts of stars of various spectral types **follow the same** spectral type **distribution** in the population.
H_A (something is going on)	Stars selected are not a simple random sample from the population of all GC stars. The observed counts of stars of various spectral types **do not follow the same** spectral type **distribution** in the population.

Example (numbers made up!)

- quantify how different the observed counts are from the expected counts
- large deviations from what would be expected based on sampling variation (chance) alone provide strong evidence for the alternative hypothesis
- this is called a **goodness-of-fi**t test since we are evaluating how well the observed data **fit** the expected distribution

Calculate the expected number of stars of each type if the sample is selected randomly:

spectral type	G,K,M	F	O	B	other	total
% in population	80.29	12.06	0.79	2.92	3.94	100
expected #	$2500\times0.8029=$	$2500\times0.1206=$	20	73	98	2500

Check row total!

Example (numbers made up!)

spectral type	G,K,M	F	O	B	other	total
% in population	80.29	12.06	0.79	2.92	3.94	100
expected #	2007	302	20	73	98	2500
observed #	1920	347	19	84	130	2500

Conditions for the chi-square test:

1. **Independence**: Sampled observations must be independent.
   - random sample/assignment
   - if sampling without replacement, n < 10% of population
   - each case only contributes to one cell in the table
2. **Sample size**: Each particular scenario (i.e. cell) must have at least 5 expected cases.

Anatomy of a test statistic

General form of a test statistics

Out[ ]//TraditionalForm=

$$\frac{\text{point estimate} - \text{null value}}{\text{SE of point estimate}}$$

1. identifying the difference between a point estimate and an expected value if the null hypothesis were true
2. standardizing that difference using the standard error of the point estimate

Chi-square statistic

When dealing with counts and investigating how far the observed counts are from the expected counts, we use a new test statistic called the chi-square (χ^2) statistic

Out[ ]=

χ^2 statistic	$\chi^2 = \sum_{i=1}^{k} \frac{(O-E)^2}{E}$	O: observed E: expected k: number of cells

By squaring the deviations we treat positive and negative values equally and there is a higher penalty for greater deviations.

degrees of freedom

To determine if the calculated χ^2 statistic is considered unusually high or not we need to first describe its distribution. The chi-square distribution has just one parameter:

degrees of freedom (df): influences the shape, center and spread.

χ^2 degrees of freedom for a GOF test $\boxed{\mathrm{df} = k - 1}$ k: number of cells

χ^2 distribution

If $Z_1, \ldots, Z_k$ are independent, standard normal random variables, then the sum of their squares

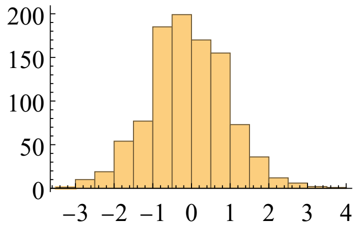
Out[∘]//TraditionalForm=

$$Q = \sum_{i=1}^{k} Z_i^2$$

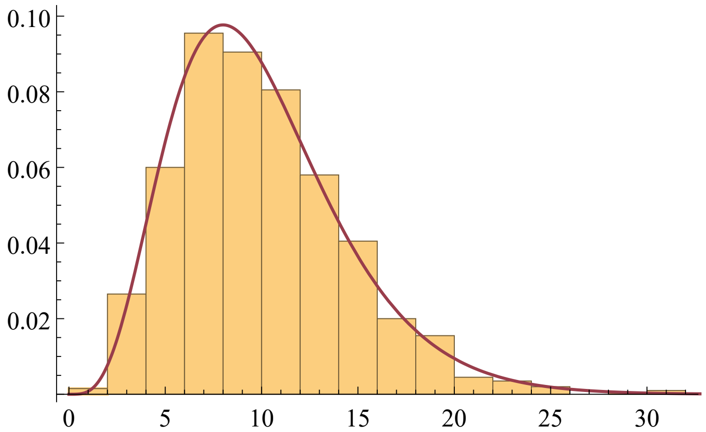
is distributed according to the chi-squared distribution with k degrees of freedom, denoted $Q \approx \chi_k^2$.

In[∘]:=
```
sample = RandomVariate[NormalDistribution[0, 1], 1000];
Histogram[sample, ImageSize → Small]
```

Out[∘]=



Out[∘]=

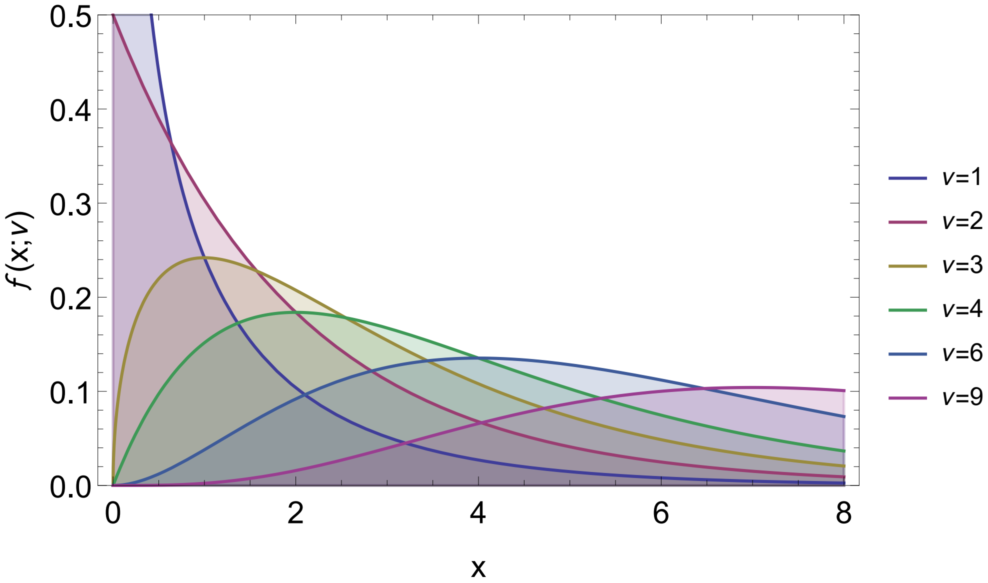


χ^2 distribution

PDF - probability density function

Out[ ]//TraditionalForm=

$$f(x\,;\,\nu) = \frac{1}{2^{\frac{\nu}{2}}\,\Gamma\left(\frac{\nu}{2}\right)}\,x^{\frac{\nu}{2}-1}\,e^{-\frac{x}{2}}$$

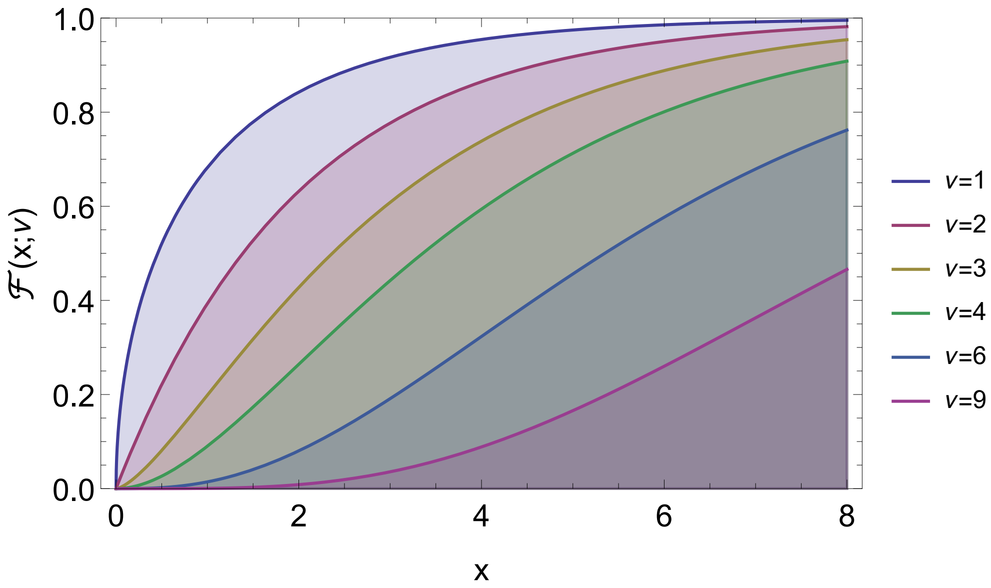


χ^2 distribution

CDF -cumulative distribution function

Out[ ]=

$$\boxed{\mathcal{F}(x;\nu) = \frac{1}{\Gamma\left(\frac{\nu}{2}\right)}\,\gamma\left(\frac{\nu}{2},\frac{x}{2}\right)}$$ with the incomplete Gamma function $$\boxed{\gamma(s,x) = \int_0^x t^{s-1}\,e^{-t}\,dt}$$



Example (numbers made up!)

spectral type	G,K,M	F	O	B	other	total
% in population	80.29	12.06	0.79	2.92	3.94	100
expected #	2007	302	20	73	98	2500
observed #	1920	347	19	84	130	2500

H_0 Stars selected are a simple random sample from the population of all GC stars. The observed counts of stars of various spectral types **follow the same** spectral type **distribution** in the population.

H_A Stars selected are not a simple random sample from the population of all GC stars. The observed counts of stars of various spectral types do not **follow the same** spectral type **distribution** in the population.

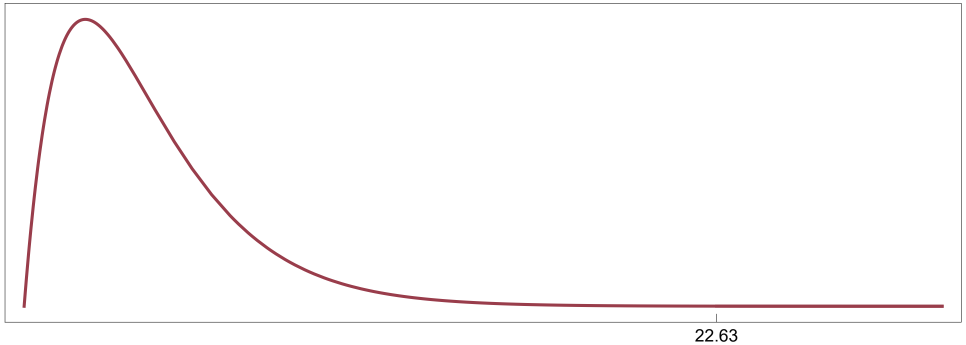
$\chi^2 = (1920-2007)^2/2007 + (347-302)^2/302 + (19-20)^2/20 + (84-73)^2/73 + (130-98)^2/98 = 22.63$

$\text{df} = k - 1 = 5 - 1 = 4$

Example (numbers made up!)

p-value

p-value for a chi-square test is defined as the tail area above the calculated test statistic. Because the test statistic is always positive, and a higher test statistic means a higher deviation from the null hypothesis.




```
1 - CDF[ChiSquareDistribution[4], 22.63]
```

```
0.000150104
```

The χ^2 GOF test shows that there is convincing evidence that the 2500 sampled stars are not an independent random sample of the GC population.

Chi-square (χ^2) independence test

Condition: at least two categorical variables with more than 2 levels.

Example (numbers made up!)

From a sample of stars with and without exoplanets we try to figure out whether their planet formation is dependent of the spectral type of the star. Let’s assume we have the following data:

	M star	G star	F star	total
no planets	81	103	147	331
w. planets	359	326	277	962
total	440	429	424	1293

Does there appear to be a relationship between spectral type and formation of exoplanets?

Example (numbers made up!)

Hypotheses:

H_0 (nothing going on):	spectral type and formation of exoplanets are **independent**. Formation of exoplanets does not vary by spectral type.
H_A (something going on):	spectral type and formation of exoplanets are **dependent**. Formation of exoplanets does vary by spectral type.

Example (numbers made up!)

- quantify how different the observed counts are from the expected counts
- large deviations from what would be expected based on sampling variation (chance) alone provide strong evidence for the alternative hypothesis
- this is called an **independence** test since we are evaluating the relationship between two categorical variables.

Chi-square test of independence

Out[ ]=

χ^2 test of independenc	$\chi^2 = \sum_{i=1}^{k} \frac{(O-E)^2}{E}$	O: observed E: e
		k: number of cells
	$df = (R-1)(C-1)$	R: number of ro
		C: number of co

Conditions for the chi-square test:

1. **Independence**: Sampled observations must be independent.
   - random sample/assignment
   - if sampling without replacement, n < 10% of population
   - each case only contributes to one cell in the table
2. **Sample size**: Each particular scenario (i.e. cell) must have at least 5 expected cases.

Example (numbers made up!)

	F star	G star	M star	total
no planets	81	103	147	331
w. planets	359	326	277	962
total	440	429	424	1293

What is the overall rate of stars without planets: 331/1293 = 0.256

If in fact planet formation and spectral stellar type are independent, i.e. if in fact H_0 is true, how many of the M type stars would we expect to have no planet? (This is our model.)

M type: $440 \times 0.256 \approx 113$

G type: $429 \times 0.256 \approx 110$

F type: $424 \times 0.256 \approx 108$

	M star	G star	F star	total
no planets	81 (113)	103 (110)	147 (108)	331
w. planets	359 (327)	326 (319)	277 (316)	962
total	440	429	424	1293

Check that there no rounding errors by comparing to the row and column totals.

Example (numbers made up!)

	M star	G star	F star	total
no planets	81 (113)	103 (110)	147 (108)	331
w. planets	359 (327)	326 (319)	277 (316)	962
total	440	429	424	1293

Test the hypothesis that stellar spectral type and planet formation are associated at the 5% significance level.

$\chi^2 = \frac{(81-113)^2}{113} + (103-110)^2/110 + (147-108)^2/108 + (359-327)^2/327 + (326-319)^2/319 + (277-316)^2/316 = 31.68$

$\mathrm{df} = (2-1)\times(3-1) = 1\times 2 = 2$

```
1 - CDF[ChiSquareDistribution[2], 31.68]
```

1.32061×10^{-7}

Example (numbers made up!)

With a small p value, we reject the null hypothesis in favor of the alternative, which means that these data provide convincing evidence that stellar spectral type and planet formation are associated.

Can we conclude that the spectral type of a star influences the planet formation process?

NO! The type of analysis that we conducted here is simply not sufficient to deduce a causal relationship. It could be due to other factors, e.g. age of the system.

Init