Testing Hypotheses and Chi–Square Tests

The term **Parameter** refers to [unknown] values pertaining to the Population like True Mean and True Proportion that *need* to be estimated...since it's not feasible or practical to calculate this for large populations! At best, we might surmise their value based some earlier study: *Hmm, we believe that the True Mean annual male incomes is $55000 now because that's what a large survey of employers said 4 years ago and adjusting for an annual 2.5% inflation...*or it may simply be a goal to aim for: *Hmm, we hope the true proportion of defectives in a manufacturing process, P = 0.01%.* Again, the True Proportion of defective TVs or True Mean Incomes of CA residents, are Parameters of interest to a TV manufacturer and to economists.

The term **Statistic / Estimate** (both terms are used interchangeably) refers to values related to the Sample, like Sample Proportion *used* to estimate or verify the value of the Parameter. In the above instance, the Sample Proportion of defective TVs – calculated from a random sample of TV sets - would be the Estimate [or Statistic]; likewise, the sample mean incomes – obtained from a random sample – would constitute the statistic.

Since it's impossible to actually determine the value of the Parameter, we use data derived from a [random] sample -- or a survey -- to *estimate* it. The Statistic / Estimate obtained from the sample are used to draw conclusions about the [unknown] Parameter(s).

After all, it is not practical to determine the *actual* proportion of defectives amongst millions of TV sets, nor would it be feasible to inquire of the starting salaries of ALL administrative assistants. We may merely have a hypothesis [i.e. a guess / claim] about the parameter value [the true proportion of ALL defectives is, say, 8%; the true mean starting salaries for ALL administrative assistants is $35,000]. Of course, the hypothesized value is seldom the "true" population value; it is merely a hunch *or* a result from previous research.

Key Logic of Statistical Inference
We use the statistic to *estimate* the parameter and *draw inferences* i.e. make conclusions, about the value of the parameter since it is not feasible to measure the parameter itself.

After all, it is not practical to inquire of the salaries of ALL typists or determine the ACTUAL proportion of defectives amongst millions of items! We may merely have a hypothesis [i.e. a guess / claim / theory] about the parameter value [the true mean salaries for ALL typists is $35,000; the true proportion of ALL defectives is 2%]. Of course, the hypothesized value is seldom the "true" population value; it is merely a hunch *or* the result from previous research.

We may wish to *test* this Hypothesis – called the **Null Hypothesis**, designated as Ho – a statement of "no change", that the *status quo* still holds, that there is no difference between the current and a prior situation.

We do this by developing an **Alternate Hypothesis** – referred to as Ha or H1 – a claim stating that a change *has* occurred in the parameter's value: perhaps, it's *higher* than or *lower* than before, or simply, *different* from that which was earlier speculated.

But how *do* we examine the "truth" of the [Null] hypothesis? We assess the validity of the original claim by taking a random sample from the population, computing the relevant estimate [for instance, the sample mean salary from a random sample of 45 typist salaries; the sample proportion of defective items from a random sample of 50 items] and determine how *extreme* or *rare* the observed outcomes would be *if* the claim in the Null Hypothesis *were* indeed true.

Based on this, we decide if the *original* parameter value is still tenable or has changed significantly. In other words, we draw a conclusion i.e. make an Inference about the true value of the Parameter – is it higher than we speculated? Is it lower? Or is it as we initially claimed? – based on the Statistic we obtain from a Sample.

This whole process:

- Formulating a Hypothesis regarding the Parameter,
- Taking a Random Sample and Calculating the [corresponding] Statistic,
- Examining *if the initial claim is true,* how unusual is it to get a result as extreme as the Statistic (which is calculating the P-value!)...and finally,
- Drawing an Inference about the Parameter based on the P-value

is called performing a Test of Significance.

Statistical Significance
An outcome is considered to be **statistically significant** [or simply, *significant*] *if* it is so "extreme" that its occurrence cannot be attributed to random chance or natural sampling variations. For example, getting ninety-nine 1s and one 2 in 100 rolls of a die is so extreme that it would never occur by chance!

Alternately, an observed result is statistically significant if it is "rare" i.e. if the initial claim in Ho were true, then we would *not* get the outcome merely due to sampling or chance variations! [think of it this way: if the statistic *is* rare, then it is significantly "different" from our claim in Ho!]

If an outcome is *not* rare, then it is **statistically *not* significant** [think of it this way: if the statistic is not rare, then it is not significantly "different" from the hypothesis in Ho!]

P-Value: simply gives the probability of getting a result as extreme as that observed if Ho were indeed True.

Significance Level, α: is just the % *below* which an outcome is regarded as rare or statistically significant. In *most* situations, α is 5% or 1%.

Chi-Square Test of Goodness of Fit

1. A "model" is stated regarding multiple proportions, with ONE sample drawn and classified into multiple categories according to some characteristic to "verify" it. **<––––––––––– understand this well!**
2. We are interested in: **<––––––––––– understand this well!**
Null Hypotheses, Ho: the sample data "fit" [is consistent with] the claim of [...].
Alternate Hypotheses, H1: the sample data do not "fit" [is *not* consistent with] the claim of [...].
OR

The multiple proportions *may* be the same [in the case of a "uniform" distribution"] so that
Ho: P1 = P2 = P3 = ...Pn = *p*% **or** the distribution is uniform
H1: At least 1 proportion in Ho is different **or** the distribution is *not* uniform.

3. The conditions to check for are:

a) the data are a random sample OR (at the very least) may be regarded as representative of the population

b) the outcomes or responses are independent of each other

c) At least 80% i.e. 4 of every 5 categories, the expected frequencies are ≥ 5.

4. CALCULATIONS: Compute the Expected frequencies by first finding the Total of the Observed Frequencies [this is the Grand Total]. Then, use the hypothesized % [stated / implied in Ho] to find the individual Expected Frequencies based on the Grand Total.

Perform Calculator Computations:

- Enter **L1** ~ Observed Frequencies, **O**; **L2** ~ Expected Frequencies, **E**
- L3 = $(O - E)^2/E$ ~ $\mathbf{(L1 - L2)^2/L2}$ **<--------------------- Use ( )!**
- Calculate **Test Statistic** $\boldsymbol{\chi}^2{}_{(n-1)\,df} = \boldsymbol{\Sigma(O - E)^2/E}$ where *n* is the number of categories in the question or table. **Show work** *re* use of the formula**.**
  Note: Use STAT → CALC → 1–Var Stats L3 to find the sum of L3 values.
- The $\boldsymbol{\chi}^2$ test-statistic for the Goodness of Fit test is associated with *n* – 1 degrees of freedom.
- Alternately, use the **STAT → TESTS →** $\boldsymbol{\chi^2}$**-GOF-Test** command with Observed values in L1 and Expected Frequencies in L2.
- Determine the Critical Value, χ^{2*} using Chi-Square tables [provided] and decide whether to reject Ho: examine the intersection of *n* – 1 degrees of freedom and the given significance level, α [usually 1% or 5%]
- **Making your decision Method 1:**
  If the computed value of $\boldsymbol{\chi}^2 \geq$ Critical Value, $\boldsymbol{\chi}^{2*}$ we reject the Null Hypotheses and find in favour of the Alternate Hypotheses i.e. we find the results to be statistically significant.
  If the computed value of $\boldsymbol{\chi}^2 \leq$ Critical Value, $\boldsymbol{\chi}^{2*}$ we do not reject the Null Hypotheses since we did not find sufficient evidence in favour of the Alternate Hypotheses i.e. we find the results to be statistically *not* significant.
- **Making your decision Method 2**
  Find the P–value, P = P($\boldsymbol{\chi}^2 > X^2$–value) = VARS → $\boldsymbol{\chi}^2$cdf(X^2–value, 9999, *df*) where *degrees of freedom, df* ~ *n* – 1 with *n* being the Number of Categories
  If the P-value, P < Significance Level, α (usually 1% or 5%), we reject the Null Hypotheses and find in favour of the Alternate Hypotheses i.e. we find the results to be statistically significant.

  If the P-value, P > Significance Level, α (usually 1% or 5%), we do not reject the Null Hypotheses since we did not find sufficient evidence in favour of the Alternate Hypotheses i.e. we find the results to be statistically *not* significant.

5. Write a Conclusion.

Problem. Census data for New York City indicate that 29.2% of the under-18 population is white, 28.2% black, 31.5% Latino, 9.1% Asian and 2%, other ethnicities. The New York American Civil Liberties Union points out that of 26,181 police officers, 64.8% are white, 14.5% black, 19.1% Hispanic and 1.4% Asian. Do the police officers reflect the ethnic composition of the city's youth?

Test an appropriate hypotheses and write your conclusion.

Solution.

Hypotheses:

Ho: The NY city police officers reflect the ethnic composition of the city's youth **OR** The ethnic distribution of the NY city cops is consistent with / matches that of the city's youth. **<–––– this is better!**

H1: The NY city police officers **do NOT** reflect the ethnic composition of the city's youth **OR** The ethnic distribution of the NY city cops is **NOT** consistent with / **does NOT** match that of the city's youth. **<–––– this is better!**

Expected Frequencies, under Ho [based on 26181 cops]:

	Observed Frequencies, O	**Expected Frequencies, E**
W	16965	29.2% of 26181 = 7644.85
B	3796	28.2% of 26181 = 7383.04
L	5001	31.5% of 26181 = 8247.02
A	367	9.1% of 26181 = 2382.47
O	52	2% of 26181 = 523.62
Total	26181	26181

Conditions

Assume that 26181 cops are representative of the ethnic distribution of cops of NY city [in general i.e. historically] and ethnicity are independent of each other. Since all expected frequencies are > 5, we can proceed with the Chi–Square Test of Goodness of Fit.

Clarification: *All* expected frequencies need *not* exceed 5. The Chi–Square Tests are valid even if *most* of them do; specifically, up to 20% [i.e. 1/5th] of the cells *can* be < 5.

Calculations

Under Ho, X^2 (4) = $\Sigma(O - E)^2/E$ = **Show how the formula is used!** $(16965 - 7644.85)^2/ 7644.85 + ...(52 - 523.62)^2/523.62 = 16{,}500$

Sketch figure, label and shade showing: $X^{2*} = 9.488$ and $X^2 = 16{,}500$

P–value, P = $P(X^2 > 16{,}500) \approx 0\%$.

Conclusion

Our P–value of ≈ 0% indicates that if indeed the NY city police officers reflect the ethnic composition of the city's youth **OR** The ethnic distribution of the NY city cops is consistent with / matches that of the city's youth, we'd get a result as extreme as that observed, practically NEVER! Therefore, since P–value ≈ 0 < α = 5%, the observed differences are indeed statistically significant. We reject Ho at the 5% significant level concluding that the NY city police officers do *not* reflect the ethnic composition of the city's youth **OR** the ethnic distribution of the NY city cops is *not* consistent with / matches that of the city's youth.

Problem. Offspring of certain fruit flies may have yellow or ebony bodies and normal wings or short wings. Genetic theory predicts that these traits will appear in the ratio 9:3:3:1 (9 yellow, normal; 3 yellow, short; 3 ebony, normal; and 1 ebony, short). A researcher checks 100 such flies and finds the distribution of the traits to be 59, 20, 11 and 10, respectively. Are the results the researcher observed consistent with the theoretical distribution predicted by the genetic model?

Solution.

Hypotheses: Ho: The [sample] data is consistent with the hypothesized model of 9:3:3:1 **or** P1 = 9/16, P2 = 3/16, P3 = 3/16 and P4 = 1/16 [**Note:** 9:3:3:1 ~ there were some *multiple* of 16 fruit–flies in all!]

H1: The [sample] data is *not* consistent with the hypothesized model of 9:3:3:1 **or** At least 1 proportion in Ho is different.

Expected Frequencies Under Ho [based on 100 flies]

	Observed Frequencies, O	Expected Frequencies, E
YN	59	9/16·100 = 56.25
YS	20	3/16·100 = 18.75
EN	11	3/16·100 = 18.75
EN	10	1/16·100 = 6.25
Total	100	100

Note: 9:3:3:1 ~ there were 16 fruit–flies in all!

Conditions

Assume that observations / observed distribution of traits are representative of ALL results [in general] and independent of each other. Since all expected frequencies are > 5, we can proceed with the Chi–Square Test of Goodness of Fit.

Clarification: *All* expected frequencies need *not* exceed 5. The Chi–Square Tests are valid even if *most* of them do; specifically, up to 20% [i.e. 1/5th] of the cells *can* be < 5.

Calculations

Under Ho, X^2 (3) = $\Sigma(O - E)^2/E$ = **Show how the formula is used!** $(59 - 56.25)^2/56.25 + ...(10 - 6.25)^2/6.25 \approx 5.671$

Sketch figure, label and shade showing: $X^{2*} = 815$ [**obtained from Chi-Square tables with α = 5%]** and $X^2 = 5.671$

P–value, P = $P(X^2 > 5.671) \approx 12.88\%$.

Conclusion

Our P–value of ≈ 12.88% indicates if indeed the model of 9:3:3:1 is valid, then we'd get a result as extreme as that observed, in 12.88% of all experiments! Therefore, since P–value ≈ 12.88% > α = 5%, the observed differences are statistically *not* significant, attributable to natural sampling variations. We cannot reject Ho at the 5% significant level concluding that we didn't find evidence that the [sample] data was inconsistent with the hypothesized model of 9:3:3:1.

Problem. Interferons are proteins produced naturally by the human body that help fight infections and regulate the immune system. A drug developed from interferons, called Avonex, is now available for treating patients with multiple schlerosis (MS). In a clinic study, 85 MS patients received weekly injections of Avonex over a 2–year period. The number of exacerbations (i.e. flare–ups of symptoms) was recorded for each patient [Source: *Biogen Inc., 1997*].

Number of Exacerbations	Number of Patients
0	32
1	26
2	15
3	6
4 or more	6
Total	**85**

For MS patients who take a placebo (no drug) over a similar two–year period, it is known from previous studies that 26% will experience no exacerbations, 30% one exacerbations, 11% two exacerbations, 14% three exacerbations, and 19% four or more exacerbations. Conduct a test to determine whether the distribution of exacerbations of MS patients who take Avonex differs from the percentages reported for placebo patients using α = 5%. **Show ALL steps.**

Solution.

Hypotheses

Ho: The distribution of exacerbations of MS patients who take Avonex does **not** differ from the percentages reported for placebo patients [**Power Tip!** Get the phrasing ***from the question, <u>if possible</u>***!] **or** There is no significant difference in the proportion of MS patients that experienced exacerbations between the Avonex and Placebo groups for each Number of Exacerbation

H1: The distribution of exacerbations of MS patients who take Avonex **differs** from the percentages reported for placebo patients **or** There **is** a significant difference in the proportion of MS patients that experienced exacerbations between the Avonex and Placebo groups for each Number of Exacerbation

Expected Frequencies, under Ho [based on 85 exacerbations]

Number of Exacerbations	Observed Frequency (O)	Expected Frequency (E)
0	32	26% of 85 = 22.1
1	26	30% of 85 = 25.5
2	15	11% of 85 = 9.35
3	6	14% of 85 = 11.9
4 or more	6	19% of 85 = 16.15
Total	**85**	**85**

Conditions

Assume that the patients were randomly assigned to the Avonex and Placebo groups, and assume that the Number of Exacerbations are independent of each patient, between and within the 2 groups in a carefully designed controlled experiment; since all expected frequencies are> 5, we can proceed with the Chi–Square Test of Goodness of Fit.

Calculations

Under Ho, X^2(4 *df*) = Σ(O – E)2/E = **Show how the formula is used!** $(32 - 22.1)^2/22.1 + ...(6 - 16.15)^2/16.15 = 17.1631$

Sketch figure, label $X^2 = 17.1631$ and $X^{2*} = 9.49$ [**obtained from Chi-Square tables with α = 5%]** and shade.

P–value, $P = P(X^2 > 17.1631) \approx 0\%$

Conclusion

Our P–value of ≈ 0% indicates that if indeed the distribution of exacerbations of MS patients who take Avonex did not differ from the percentages reported for placebo patients **or** there was no significant difference in the proportion of MS patients that experienced exacerbations between the Avonex and Placebo groups for each Number of Exacerbation, we'd get a result as extreme as that observed, um, practically NEVER! Therefore, since P–value ≈ 0 < α = 5%, the observed differences are indeed statistically significant. We reject Ho at the 5% significant level concluding that we did find evidence that the distribution of exacerbations of MS patients who take Avonex **differs** from the percentages reported for placebo patients **or** There **is** a significant difference in the proportion of MS patients that experienced exacerbations between the Avonex and Placebo groups for each Number of Exacerbation.

Problem. According to the March 2000 Current Population Survey, the marital status distribution of the US adult population is as: Never married: 28.1%; Married: 56.3%; Widowed: 6.4% and Divorced: 9.2%. A random sample of 500 US adult males, aged 25-29 years old, yielded the following frequency distribution: Never married: 260; Married: 220; Widowed: 0; and Divorced: 20. Perform a Goodness of Fit Test to determine if the marital status distribution of US males 25-29 years old differs from that of the US adult population.

Solution.

Hypotheses: Ho: The distribution of marital status for 25–29 year old US males is consistent with that of the population.

H1: The distribution of marital status for 25–29 year old US males is *not* consistent with that of the population.

Expected Frequencies, under Ho: [based on the 500 males]:

Marital Status	Observed Frequencies	Expected Frequencies
Never Married	260	140.5
Married	220	281.5
Widowed	0	32
Divorced	20	46
Total	500	500

Conditions
Since sample of 500 US males is given to be SRS, assume marital status to be representative of all US 25–29 males and to be independent of each other (within and between categories); as all expected frequencies are> 5, we can proceed with the Chi–Square Test of Goodness of Fit.
Calculations

Under Ho, X^2(3 *df*) = Σ(O – E)2/E = **Show how the formula is used!** $(260 – 1405.)^2/140.5$ + ...$(20 – 46)^2/46$ = 161.77

Sketch figure, label X^2 = 161.77 and X^{2*} = 7.815 [**obtained from Chi-Square tables with α = 5%]** and shade.

P–value, P = P(X^2 > 161.77) ≈ 0%.
Conclusion
If the distribution of marital status for 25–29 year old US males is consistent with that of the population, our P–value of 0% indicates that we'd get a result as extreme as that observed, um, practically NEVER! Therefore, since P–value ≈ 0% < α = 5%, the observed differences are indeed statistically significant. We reject Ho at the 5% significant level and conclude that the distribution of marital status for 25–29 year old US males is **not** consistent with that of the population.

Problem. An article about the CA lottery gave the following information about the age distribution of adults in CA: 35% between 18-34 years old; 51% between 35-64 years ols; and 14% greater than 65 years old. The article also gave the age distribution of those that purchase lottery tickets: 36 between 18-34 years old; 130 between 35-64 years ols; and 34, greater than 65 years old. Suppose that the data resulted from a random sample of 200 lottery ticker purchasers. Is it reasonable to conclude that one or more of these age-groups buy a disproportionate number if lottery tickets?

Solution.

Hypotheses: Ho: There is no significant difference between the distribution of CA adults and lottery players **or** the distribution of CA adults and lottery players is *not significantly* different **or** none of the age–goups purchases a disproportionate amount of lottery tickets [***from the Q!*]** **or** P1*l* = 35%, P2*l* = 51%, P3*l* = 14%

H1: There is a significant difference between the distribution of CA adults and lottery players **or** the distribution of CA adults and lottery players is significantly different **or** at least 1 age–group purchases a disproportionate amount of lottery tickets **or** at least 1 proportion in Ho is different.

Expected Frequencies, under Ho [based on 200 adults]

Age	**O**	**E**
18–34	36	35% of 200 = 70
35–64	130	51% of 200 = 102
65 and older	34	14% of 200 = 28
Total	200	200

Conditions
Given that 200 lottery ticket purchasers were randomly selected, assume purchase behavior to be representative of their respective populations and independent of each other (within and between groups); since expected frequencies > 5, we can proceed with the Chi–Square Goodness of Fit Test.

Calculations
Under Ho, X^2(2 *df*) = Σ (O – E)^2/E = **Show how the formula is used!** (36 – 70)^2/70 + ...(34 – 28)^2/28 = 25.48

P–value, P = P(X^2 > 25.48) ≈ 0%

Sketch figure, label X^2 = 25.48 and X^{2*} = 5.99 [**obtained from Chi-Square tables with α = 5%]** and shade!

Conclusion
Our P–value of ≈ 0% shows that if indeed there was no significant difference between the distribution of CA adults and lottery players **or** the distribution of CA adults and lottery players was *not significantly* different **or** none of the age–goups purchased a disproportionate amount of lottery tickets **or** P1*l* = 35%, P2*l* = 51%, P3*l* = 14%, then we'd get results as extreme as those observed...practically never. Since P–value ≈ 0% < α = 5%, we find the results statistically significant, not attributable to sampling variations. We reject Ho at the 5% significance level, and conclude that we did find evidence that there is a significant difference between the distribution of CA adults and lottery players **or** the distribution of CA adults and lottery players is significantly different **or** at least 1 age–group purchases a disproportionate amount of lottery tickets **or** at least 1 proportion in Ho is different.