Research Methods

Robert E Simpson

Singapore University of Technology & Design

robert_simpson@sutd.edu.sg

October 23, 2018

Objectives

You will be able to:

- Use χ^2 and F-tables
- Understand that data in the bins of a histogram is distributed according to Poisson statistics, and therefore the vairance of the data in the bin is equal to the mean value of the data
- Calculate the appropriate degrees of freedom for χ^2 and F-statistic tests
- Statistically check whether two variances are equal
- Calculate the confidence limits within which one would expect the population variance based on a sample variance computed from a sample of random variables
- Describe how least squares fits are performed
- Compute the confidence that a function fits measured data
- Analytically calculate the most likely parameters of a fitting function for a given data set using Maximum Likelihood Estimation
- Use the χ^2 distribution to ascertain whether a dataset *significantly* differs from a distribution
- Use the F-distribution to state with a confidence whether the two variances are *significantly* different (and therefore state whether the samples come from the same population).

Tests of variance

Previously we used a sample mean, $\bar{x}$, to estimate the interval of values within which we expect the true population mean, μ, to exist. We will now use a similar approach for estimating the variance of the population distribution from a sample distribution.

Confidence Intervals on the Variance

Previously we predicted the confidence limits for the true mean, but we can also define confidence limits for the variance.

- Assuming Normally distributed random vairables
- Let $X_1, \ldots, X_n$, be a random sample with mean μ, population variance σ^2, and sample variance, S^2, then the random variable

$$\boxed{X^2 = \frac{(n-1)S^2}{\sigma^2}}$$

  has a χ^2 distribution with $n-1$ degrees of freedom.
- With this we can find the range of X values within which we would expect the the true variance σ based on a sample with variance, S.[1]

[1] X^2 is the proportion of the sample variance relative to the population variance.

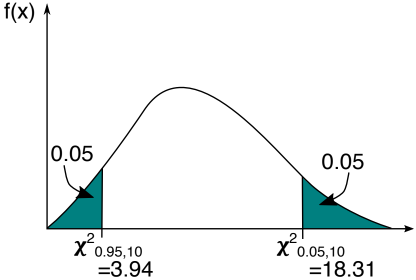
Confidence Intervals on the Variance

$$P(\chi^2_{1-\alpha/2} \le X^2 \le \chi^2_{\alpha/2}) = 1 - \alpha$$

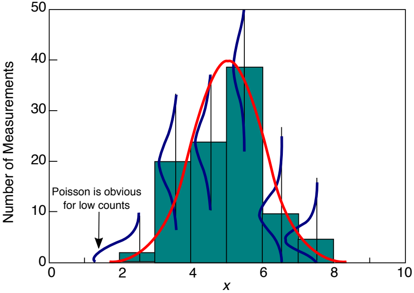
$$P(\chi^2_{1-\alpha/2} \le \frac{(n-1)S^2}{\sigma^2} \le \chi^2_{\alpha/2}) = 1 - \alpha$$

Thus:

$$\boxed{P\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2}} \le \sigma^2 \le \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}\right) = 1 - \alpha} \qquad (1)$$



The χ^2 Statistic



Histogram of a Gaussian distribution with $\mu = 5$ and $\sigma = 1$, and $N = 100$. The parent PDF is shown by the red cuve. The blue curves show the Poisson distribution of events in each bin, based on the same data.

Considering a histogram of data, χ^2 is a statistic that characterises the dispersion of the observed frequencies, $h(x_i)$, from the expected frequencies, $NP(x_i)$.

$$\boxed{\chi^2 = \sum_{i=1}^{n} \frac{[h(x_i) - NP(x_i)]^2}{\sigma_i(h)^2}} \tag{2}$$

The χ^2 Statistic

When we make a histogram, we are essentially doing a counting experiment. I.e. counting the number of occurrences within some range. Hence Poisson Statistics are applicable. Although, the distribution of frequencies maybe any function, the probability function for the spreads of measurements of each frequency are Poisson distributions.

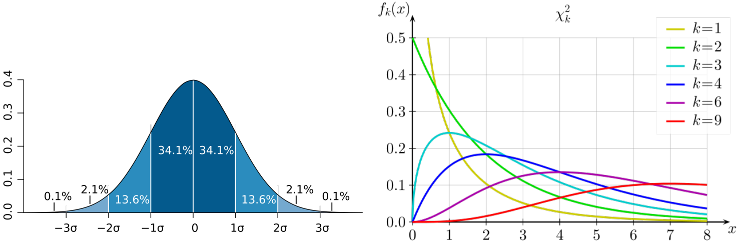
Hence: $\sigma_i(h) = \sqrt{NP(x_i)}$ and therefore:

$$\chi^2 = \sum_{i=1}^{n} \frac{[h(x_i) - P(x_i)]^2}{\sigma_i(h)^2} \approx \sum_{i=1}^{n} \frac{[h(x_i) - NP(x_i)]^2}{NP(x_i)}$$

More generally, χ^2 can be used to check how well an observed distribution f_{obs} is fitted by an expected distribution model f_{exp}.

$$\boxed{\chi^2 = \sum_{i=1}^{n} \frac{[f_{obs,i} - f_{exp,i}]^2}{f_{exp,i}}} \qquad (3)$$

χ^2 distribution



(a) Standard normal PDF

(b) chi squared PDF

- If we take one sample from the standard normal pdf, then there is a high probability that the value will be zero, so the χ^2 is weighted towards zero.
- If we take more than one sample, the square of the samples add together, thus the χ^2 PDF peaks at values greater than zero.
- Notice that the χ^2 PDF only has values for x greater than zero because we add x_i random variables in quadrature.

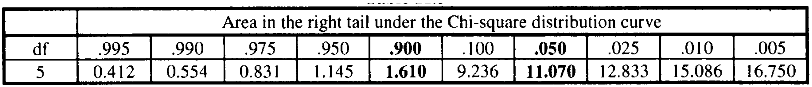
χ^2 distribution degrees of freedom

- If we take one sample from the normal distribution, then the chi squared distribution has 1 degree of freedom (d.f) and we write: χ_1^2.
- If we take 2 samples from the normal distribution, then the chi squared distribution has 2 d.f. (X_1 and X_2) and we write: $\chi_2^2 = X_1^2 + X_2^2$.
- If we take N samples from the normal distribution, then the chi-squared has N d.f, and we write: $\chi_N^2 = \sum_i^N X_i^2$.

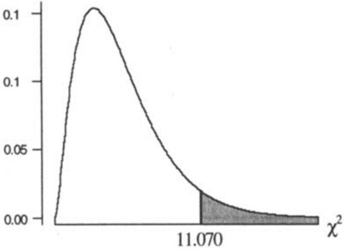
χ^2 distribution

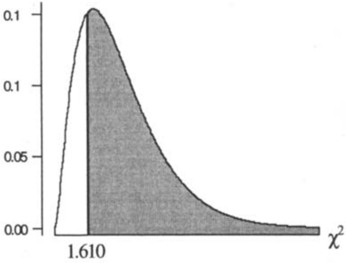
- The total area under the χ^2 PDF is one.
- Has values for $x > 0$
- Is always skewed compared to a Normal PDF
- Increasing the number of degrees of freedom (d.f.) makes it more like a normal distribution
- The mean average is equal to the number of d.f. and the variance is $2 \times d.f.$
- When $d.f. > 3$, the peak of occurs at $d.f. - 2$, this is known at the *mode* or *model* of the distribution
- There is a set of PDFs with different values of d.f.

χ^2 distributioon



	Area in the right tail under the Chi-square distribution curve									
df	.995	.990	.975	.950	**.900**	.100	**.050**	.025	.010	.005
5	0.412	0.554	0.831	1.145	**1.610**	9.236	**11.070**	12.833	15.086	16.750





Checking if two variances are equal

When we perform a t-test on two means from different samples (2 sample test), it is presumed that the variance of the two tests are equal. It is useful to have a method to test whether the variances of the two samples are in fact equal.

let σ_1 and σ_2 be the standard deviations of the two populations, and let s_1 and s_2 be the sample standard deviations. If $\sigma_1 = \sigma_2$, then the random variable:

$$F = \frac{s_1^2}{s_2^2} \tag{4}$$

has a F-distribution with d.f.$=(v_1, v_2)$ where $v_1 = n_1 - 1$ and $v_2 = n_2 - 1$

Notes: The larger sample variance is always taken as the numerator. The F-distribution applied to independent samples and is sensitive to the assumption of normality (Take care).

Fisher-Snedecor F distribution

- Defined as the ratio of two independent random X^2 variables [2]
- Continuous probability distribution
- Allows comparison of two or more sample variances (like t-distribution allows comparison of two sample means)
- Described by a set of plots, where each plot is specific to a set of numbers representing the degrees of freedom of two random variables, v_1 and v_2.
- Frequently used for analysis of variance (ANOVA)

[2] Note this is X^2 and not χ^2. In ANOVA the ratio is $F = \frac{MSTr}{MSE}$

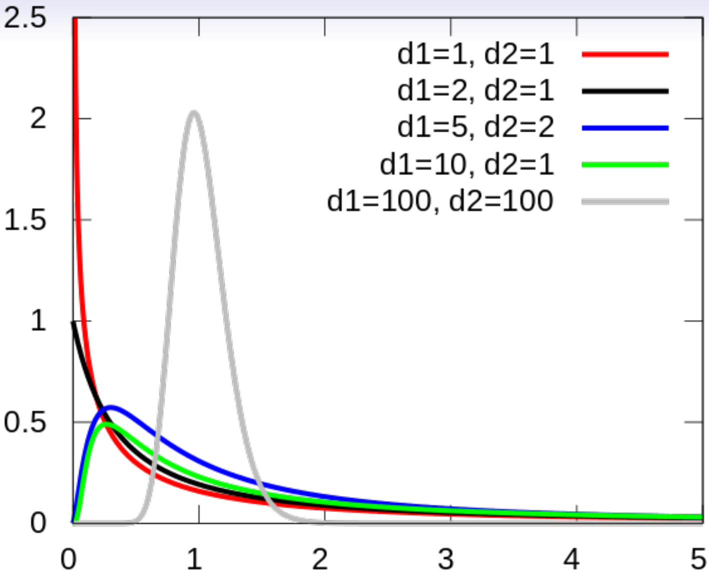
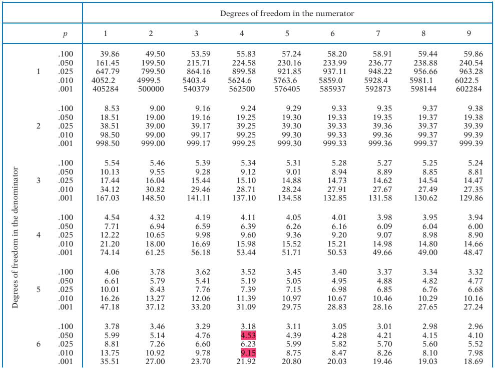


image source: wikipedia

Using F-tables Example

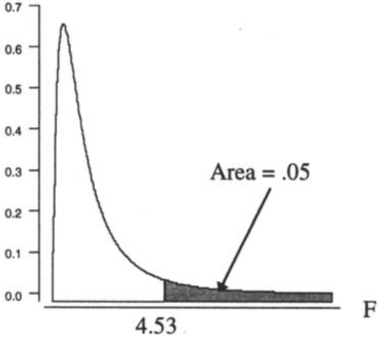
Consider the F distribution with df1 = 4 and df2 = 6. The table below shows a portion of the F distribution table. What is the critical F value for a signifficance level of 0.05 and 0.01?

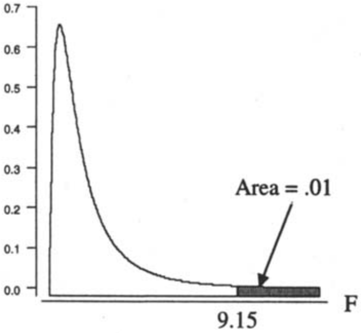


			Degrees of freedom in the numerator								
		p	1	2	3	4	5	6	7	8	9
Degrees of freedom in the denominator	1	.100	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86
		.050	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54
		.025	647.79	799.50	864.16	899.58	921.85	937.11	948.22	956.66	963.28
		.010	4052.2	4999.5	5403.4	5624.6	5763.6	5859.0	5928.4	5981.1	6022.5
		.001	405284	500000	540379	562500	576405	585937	592873	598144	602284
	2	.100	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38
		.050	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38
		.025	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39
		.010	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39
		.001	998.50	999.00	999.17	999.25	999.30	999.33	999.36	999.37	999.39
	3	.100	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24
		.050	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
		.025	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47
		.010	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35
		.001	167.03	148.50	141.11	137.10	134.58	132.85	131.58	130.62	129.86
	4	.100	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94
		.050	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
		.025	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90
		.010	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66
		.001	74.14	61.25	56.18	53.44	51.71	50.53	49.66	49.00	48.47
	5	.100	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32
		.050	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
		.025	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68
		.010	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16
		.001	47.18	37.12	33.20	31.09	29.75	28.83	28.16	27.65	27.24
	6	.100	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96
		.050	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
		.025	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52
		.010	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98
		.001	35.51	27.00	23.70	21.92	20.80	20.03	19.46	19.03	18.69

When P=0.05, F(4,6)=4.53.
When P=0.01, F(4,6)= 9.15

Using F-tables





The table indicates that the area to the right of 4.53 is .05, and the area to the right of 9.15 is 0.01

F-table notation

We use the following notation to describe F-statistics:

$$F_\alpha = (df1, df2) \tag{5}$$

where df1 is the number of degrees of freedom on the numerator of the F-statistic and df2 is on the denominator.

The value, 4.53, is represented as $F_{.05}(4,6) = 4.53$, and the value 9.15 is represented as $F_{0.01}(4,6) = 9.15$

F-statistics, left hand tail probabilities

Formula (5) may be used to find the F values for right-hand tail (as in the previous example). However, if we want to find the left-hand tail area, we can apply:

$$F_{1-\alpha}(df1, df2) = \frac{1}{F_\alpha(df2, df1)} \tag{6}$$

Thus, the symbol $F_{.95}(4,6)$ represents the value for which the area to the right of $F_{.95}(4,6)$ is 0.95 and therefore the area to the left of $F_{.95}(4,6)$ is 0.05.

Concept question 1

Find the following:

(a) $F_{.01}(7, 3)$

(b) $F_{.99}(7, 3)$

(c) The critical value for df=(6,6) and the area in the right hand tail is 0.05

Concept question 1 solutions

Find the following:

(a) $F_{.01}(7,3) = 27.67$

(b) $F_{.99}(7,3) = 1/F_{.01}(3,7) = 1/8.45 = 0.1183$

(c) The critical value is 4.28

Case problem 6.1

Decision making

A company which makes boxes wishes to determine whether their automated production line requires major service. They will base their decision on whether the weight from one box to another is significantly different from a maximum permissible population variance value of $\sigma^2 = 0.12\ kg^2$. A sample of 10 boxes is selected, and their variance is found to be $s^2 = 0.24\ kg^2$. Is this difference significant at the 95% CL?

Case problem 6.1

Decision making

Hypothesis: H_0: $\sigma = s$, and H_A: $\sigma \geq s$
Use a single sample test,

$$\chi^2 = \frac{n-1}{\sigma^2} s^2$$
$$\chi^2 = \frac{10-1}{0.12} 0.24 = 18$$

The number of degrees of freedom is $df = n - 1 = 9$.

According to the chi-squared tables for df=9, $\chi^2_{\alpha=0.05} = 16.919$, and $\chi^2_{\alpha=0.025} = 19.023$.
Therefore, for $\chi^2 = 18$, the difference between the sample and the population variance is signifficant and we reject H_0 at a 95% confidence level.
However, the difference is not significant at a 97.5% confidence level.
Therefore one might consider performing additional tests with more samples to determine if the machine actually does need servicing.

Case problem 6.2

Productivity

Productivity is thought to increase if the working environment is conditioned to meet human comfort. We want to compare the mean productivity of two workers under different conditions. However, to do a t-test we must assume equal variances in productivity of the workers (i.e., assume their output does not change from day to day). Check the validity of this assumption with the following data, which was collected under the same environment and performing the same task.
Worker A: $n_1 = 13$ days, $\bar{x}_1 = 26.3$ production units, $s_1 = 8.2$ production units.
Worker B: $n_2 = 18$ days, $\bar{x}_2 = 19.7$ production units, $s_2 = 6.0$ production units.

Case problem 6.2 Solution

Productivity

The null hypothesis is that there is no difference in the output variance of the two workers. i.e. H_0: $s_1 = s_2$ and H_A: $s_1 \neq s_2$.
We will use a two sample variance test.

$$\begin{aligned} F &= \frac{s_1^2}{s_2^2} \\ F &= \frac{8.2^2}{6^2} = 1.867 \end{aligned} \tag{7}$$

$$df_1 = n_1 - 1 = 13 - 1 = 12$$

$$df_2 = n_2 - 1 = 18 - 1 = 17$$

From the F-tables, $F_{12,17} = 2.38$ at $\alpha = 5\%$, $F_{12,17} = 1.96$ at $\alpha = 10\%$
Therefore we cannot reject H_0 at a 5% or 10% significance levels.
Therefore we must accept the null hypothesis and conclude that the variance in the output of the two workers is the same. Thus we can use a t-test to check to see if there is a statistical difference in the productivity.

Curve Fitting

- We cannot always fit a curve to go through all the points in a x vs y graph
- We often assume that points not hitting a predicted value to be due to measurement uncertainty

Least Squares Method

The objective is to find parameters of an equation that minimize the discrepency between the measured value y_i and the calculate value $y(x)$.

For a linear function, $y = mx + c$, we want to find the value of m and c that minimize discrepencies. The method for a linear example is:

- For any arbitary m and c we calculate the deviation Δy_i between y_i and the calculated value: $y(x)$.

$$\Delta y_i = y_i - y(x_i) = y_i - mx_i - c$$

- Sum the square of the deviations
- Optimise the fit by minimising the sum of square deviaitions

Method of Maximum Likelihood

The objective is to find a set of parameters for a function that are most likely to give rise to an observed set of measurements.

- The true value of a function y, y_0, depends on the true values of a set of independent variables. For simplicity we will use a linear function with true values m_0 and c_0, such that: $y_0 = m_0 x + c_0$.
- Assume that each individual y-point, y_i, is drawn from a Gaussian distribution with mean $y_0(x_i)$ and standard deviation σ.
- The probability, P_i, of making the observed measurement, y_i is then:

$$P_i = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{1}{2} \left(\frac{y_i - y_0(x_i)}{\sigma} \right)^2 \right\}$$

Method of Maximum Likelihood

- The probability of making a set of N measurements of the value y_i is the product of the probabilities for each oberservaiton:

$$L = \Pi_{i=1}^{N} P_i = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^{N} \Pi_{i=1}^{N} \exp\left[-\frac{1}{2}\sum\left(\frac{y_i - y_0(x_i)}{\sigma}\right)^2\right] \tag{8}$$

- σ_i effectively acts as a weighting factor so if σ_i is small the a large weight is given to that particular point's probability
- We can now maximise this function to find the values of x_i that are most likely to give rise to the observed set of N measurements, y_i.

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Maximum Likelihood Estimation

- Equation 8 is maximum when the argument of the exponent, $\frac{\chi^2}{2} = -\frac{1}{2}\sum\left(\frac{y_i - y_0(x_i)}{\sigma_i}\right)^2$, is minimised.[3]

$$\frac{\chi^2}{2} = -\frac{1}{2}\sum\left(\frac{y_i - y_0(x_i)}{\sigma_i}\right)^2 \tag{9}$$

$$\frac{d\chi^2}{dy_0} = \frac{d}{dy_0}\left\{\sum\left(\frac{y_i - y_0(x_i)}{\sigma_i}\right)^2\right\} = 0 \tag{10}$$

- This can be solved analytically (see slide 31 for the procedure) or numerically (problems class)
- We define χ^2 to be our Goodness of Fit.

$$\chi^2 = \sum\left(\frac{y_i - y_0(x_i)}{\sigma_i}\right)^2 \tag{11}$$

We are essentially comparing the difference between the measured and true value with the expected standard deviation, σ, which are 2 random sq. variables hence the use of χ^2. The smaller the value of χ^2, the better the fit.

[3] The minimisaiton of a linear function is gievn on P109 of Bevington. We will also do it numerically in the problems class.

Maximum Likelihood Estimation

- It can handle any type of error distribution.
- Straight Forward- easily solved by computers
- Can show a range of plausible values and for deducing confidence limits
- Can be used when there is no knowledge of the underlying distribution but needs an analysts input.

MLE Procedure

1. Define the probability distribution function
2. Define the Likelihood function, $L = \Pi_{i=1}^{n} p(x_i)$
3. Take the logarithm of L
4. Differentiate L with respect to the parameter being investigated
5. Set the differentiated function to zero, and solve.

However, this procecure does not always work. Therefore, you should also be able to find the maximum likelihood value using numerical methods (See week 6 problems class).

Case Problem 6.3

MLE

Calculate the most likely value of λ for a specifc set of Poisson random variables, $x_1, \ldots, x_n$.

Recall that the Poisson distribution function is:

$$P(x, \lambda) = \frac{\lambda^x}{x!} e^{-\lambda}$$

CP 6.3 Solution

$$L = \Pi_{i=1}^{N} P_i = \Pi_{i=1}^{N} \left(\frac{\lambda^{x_i}}{x_i!} e^{-\lambda} \right)$$

Take the natural log:

$$\ln L = \sum_{i}^{N} x_i \ln(\lambda) - \sum_{1}^{N} \ln(x_i!) - N\lambda$$

Differentiate wrt λ

$$\frac{d(\ln(L))}{d\lambda} = \sum_{i}^{N} x_i \frac{1}{\lambda} - N = 0$$

Thus the most likely value of λ is:

$$\lambda = \frac{\sum_{i}^{N} x_i}{N} = \bar{x}$$

χ^2 Goodness of fit

- Used to ascertain whether a set of samples differs *significantly* from an expected distribution
- The χ^2 test statistic is computed as:

$$\boxed{\chi_o^2 = \sum_k \frac{(f_{obs,k} - f_{exp,k})^2}{f_{exp,k}}} \tag{12}$$

Where $f_{obs,k}$ is the observed frequency, and $f_{exp,k}$ is the expected frequency at point k.

χ^2 Goodness of fit

- If the population follows the hypothesised distribution, then χ_0^2 has approximately a chi-squared distribution.
- If $\chi^2 = 0$, then f_{obs} and f_{exp} agree perfectly
- Larger χ^2 indicate larger discrepancies
- χ^2 tables are used to determine the significance for different values of d.f $= k - p - 1$. p is the number of parameters estimated by the sample statistics. (e.g. if you estimate the population mean from the sample population, then p=1, but if you know the population mean, then p=0)

χ^2 Goodness of fit procedure

i Take a random sample of size *n* from the population and create a histogram of the data

ii State the null and research hypothesis concerning the hypothesized distribution for the *k* categories (histogram bins).

iii Use the χ^2 table and the level of significance, α, to determine the rejection region.

iv Compute $\chi^2 = 0$, then f_{obs} and f_{exp}, and check to make sure all expected frequencies are 5 or more.

v State your conclusion. The null hypothesis is rejected if the computed value of the test statistic falls in the rejection region. Otherwise, the null hypothesis is not rejected.

Case Problem 6.4

Goodness of fit

The number of defects in printed ciruit boards (PCB) is hypothesised to follow a Poisson distribution. The defects are counted in a random sample of 60 circuit boards. The number of defects per a PCB are counted. Can we conclude that the number of defects is Poisson distributed?

Number of defects	Frequency
0	32
1	15
2	9
3	4

Case Problem 6.4 solution

Goodness of fit

H_0: Poisson distribution, H_A: Not Poisson distribution

$$P(x) = \frac{e^{-\mu}\mu^x}{x!}$$

Estimate the mean from the sample:

$$\mu = (32 \times 0 + 1 \times 15 + 2 \times 9 + 3 \times 4)/60 = 0.75$$

$$P(0) = \frac{e^{-0.75} \times 0.75^0}{0!} = 0.472$$
$$P(1) = \frac{e^{-0.75} \times 0.75^1}{1!} = 0.354$$
$$P(2) = \frac{e^{-0.75} \times 0.75^2}{2!} = 0.133$$
$$P(X \geq 3) = 1 - P(0) - P(1) - P(2) = 0.041$$

Case Problem 6.4 solution

Goodness of fit

Since the expected frequency will be less than 5, for 3 or more defects ($60 \times P(X \geq 3) = 2.46$), we combine the results for 2 defects and 3 or more defects.

Num of defects	Frequency	P	Expected	$(f_{obs} - f_{exp})^2 / f_{exp}$
0	32	0.472	28.32	0.48
1	15	0.354	21.24	1.83
≥ 2	13	0.174	10.44	0.62

Table: Observed and expected number of defects

$$\chi_0^2 = 0.48 + 1.83 + 0.62 = 2.93$$

Case Problem 6.4 solution

Goodness of fit

Reject H_0 if $P < 0.05$, "If P is low, reject H-oh"

- From the χ^2 tables, we look up $\chi^2 = 2.93$ for one degrees of freedom ($d.f = k - p - 1 = 3 - 1 - 1 = 1$).
- We see that $\chi^2 = 2.71$, $Prob = 0.1$, and $\chi^2 = 3.84$, p=0.05.
- Therefore 2.93, has a significance level greater than 0.05 and less than 0.1.
- We **CANNOT** reject the H_0.
- We conclude at the 95% confidence level that the data is statistically similarly distributed to a Poisson distribution.