Statistics Applied to Economics

Degree in Economics

F. Tusell

Dpto. Economía Aplicada III (Estadística y Econometría)

Curso 2012–2013

eman ta zabal zazu

Universidad del País Vasco / Euskal Herriko Unibertsitatea

Índice I

Goodness-of fit problems

- Quite common hypothesis.
  1. Do winning numbers in the Lotería Primitiva appear to come from a discreet uniform distribution over {1,2,...,49} ? *(no parameters estimated, fully especified distribution)*
  2. Does the number of dead people by horse (or mule) kick in the Prusian army follow a Poisson distribution (plausible; small probability, many people at risk). *(one parameter to be estimated)*
  3. Do intervals between accidents at work appear to follow an exponential distribution? *(one parameter to be estimated)*
- In all these cases, we have data and we want to test adequacy of a given distribution, possibly not fully especified (= some parameter has to be estimated).

Test statistic

- Break down the range of the random variable in k classes. Call O_i the number of observations in class i, $i = 1, 2, \ldots, k$.
- Call E_i the number of expected observations in class i under the null hypothesis (i.e., if the assumed distribution for the data is "true").
- Then,

$$Z = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \overset{H_o}{\sim} \chi^2_{k-p-1}$$

- k is the number of classes, p the number of parameter estimated, if any.

The gory details

- Where does this come from? Proof not trivial, distribution valid only as an approximation for "large" samples.
- How large is "large"? No class should have an expected value less than, say, 5. If it does, merge classes.
- How to choose k? Reasonably large, but keeping classes "well peopled".
- Howto choose the class boundaries? Good question.
- Usually no particular alternative: a pure significance test.
- Critical region: right tail.

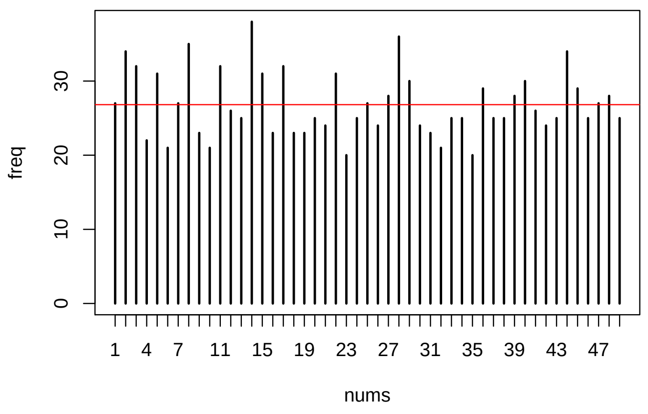
Example - I

```
> primitiva[1:3,1:8]
       Fecha Semana N1 N2 N3 N4 N5 N6
1 01/01/2009      1  4  8 12 25 34 46
2 03/01/2009      1  9 11 21 30 31 44
3 08/01/2009      2  7 17 27 28 29 44
> nums <- as.matrix(primitiva[,3:8])
> freq <- table(nums)
> sum(freq)           # How many numbers seen?
[1] 1314
> e <- sum(freq) / 49  # Expected each under H0
> e
[1] 26.81633
```

Example -II

The absolute frequencies of each number are:

```
> plot(freq)
> abline(h = e, col = "red")
```



Example -III

- Question is now to decide whether the departures from the expected number of appearances is enough to reject H_0 (="all numbers equally likely").
- We can use a χ^2-test where each "class" I is made of one number, O_i are the observed occurrences and $E_i = 26.81633$.

```
> Z <- sum((freq - e)^2/e)
> Z
[1] 33.68645
> 1 - pchisq(Z, df = 49 - 1)
[1] 0.9415792
```

- The probability in the tail is quite large; H_0 gives a very good fit and is not rejected.

Example - IV

- R has a standard function which does the same at once.

```
> result <- chisq.test(x=freq,p=rep(1/49,49))
> result
          Chi-squared test for given probabilities

data:  freq
X-squared = 33.6865, df = 48, p-value = 0.9416
```

- So, in conclusion, no evidence of "lucky" numbers.

Example - V

- If you have to do it manually, your best bet is to arrange computations in a small table.
- For instance, you might have in the case shown:

O_i	E_i	$(O_i - E_i)$	$(O_i - E_i)^2$	$(O_i - E_i)^2/E_i$
27	26.81633	0.183673	0.03373	0.001258
34	26.81633	7.183673	51.60516	1.924394
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
25	26.81633	-1.816327	3.29904	0.123023
			$Z =$	33.6865

Chi square test with estimated parameters (I)

- Data: deaths by horse kick in 200 army corps years.

Deaths	Observed cases
0	109
1	65
2	22
3	3
4	1

- Is the Poisson distribution a good model for these data?
- The hypothesis does not uniquely fix the distribution.
- The MLE of λ is:

$$\hat{\lambda} = \frac{0 \times 109 + 1 \times 65 + 2 \times 22 + 3 \times 3 + 4 \times 1}{200} = 0.61$$

Chi square test with estimated parameters (II)

- We have estimated one parameer (λ).
- We do as before, only the E_i are compute from the $\mathcal{P}(\lambda = 0.61)$ distribution.
- For instance, since

$$P(x = 0; \lambda = 0.61) = \frac{e^{-0.61}(0.61)^0}{0!} = 0.5433509$$

we would compute E_1 (the expected number of cases with 0 deaths) as: $200 \times 0.5433509 = 108.67$.

- Likewise for the remaining E_i cells.

Chi square test with estimated parameters (III)

- Now we have:

O_i	E_i	$(O_i - E_i)$	$(O_i - E_i)^2$	$\frac{(O_i-E_i)^2}{E_i}$
109	108.67017	0.32983	0.10879	0.00100
65	66.28881	-1.28881	1.66102	0.02505
22	20.21809	1.78191	3.17522	0.15704
3	4.11101	-1.11101	1.23434	0.30025
1	0.62693	0.37307	0.13918	0.22200
			$Z =$	0.70537

Chi square test with estimated parameters (IV)

- We now compare Z with a chi-square with **3** degrees of freedom ($k - p - 1 = 5 - 1 - 1 = 3$):

```
> 1 - pchisq(0.70537, df = 3)
[1] 0.8719401
```

- The tail is 0.87194; there is no reason to reject the Poisson distribution hypothesis.
- One might question the use of the test in that some classes are very sparsely populated.

Contingency table analysis

- A two-dimensional contingency table is an array which classifies observations according to two variables, one occurring in rows, the other in columns.
- Definition can be generalized to different number of dimensions
- For example, we may have:

Gender	Right-handed	Left-handed	Total
Male	43	9	52
Female	44	4	48
Total	87	13	100

- The row and column totals are referred as the *margins*.

Sampling schemes (I)

- We may fix only the total number of cases we cross-tabulate…
- …or we may fix the row margin or the column margin.
- In the first case we speak of *multinomial sampling*, in the second of *product multinomial sampling*.
- Why should we care? Marginal probabilities can only be estimated from "free" margins.

Sampling schemes (II)

- Consider the following case: we pick a sample of 1000 persons and cross-classify them according to ethnic origin and whether they suffered in the last winter from common cold. Want to test relative vulnerability.

Race	Had cold	Didn't have cold	Total
Whites	801	104	905
Non-whites	83	12	95
Total	884	116	1000

- We may estimate the proportion of whites as $905/1000 = 0.905$ and the overall prevalence of cold as 0.884

Sampling schemes (III)

- Suppose though we are sampling a population with a tiny proportion of non-whites. We might end up with a table such as:

Race	Had cold	Didn't have cold	Total
Whites	891	108	999
Non-whites	1	0	1
Total	892	108	1000

- We end up with a table in which non-whites are almost (or totally) absent.
- Non-white sample far too small to investigate the matter of interest.

Sampling schemes (IV)

- What we need instead is to sample both races separately, say 500 each:

Race	Had cold	Didn't have cold	Total
Whites	398	102	500
Non-whites	403	97	500
Total	801	199	1000

- *Then* we are assured to have enough observations in each group.
- Marginal totals do not estimate anything now: the row totals are fixed by design.

Sampling schemes (V)

- If we fix only the total, we are sampling **one** population. The hypothesis of interest is *independence* in that population.
- If we fix the row totals, we are in effect sampling **two** populations. The hypothesis of interest is *homogeneity* of both populations with respect to the character coded in columns.
- Both hypothesis are tested conditional on the margins, and the results are exactly the same for a given table, no matter how it was sampled.
- Why conditionally on the margins? It is the distribution of counts inside the table what is indicative of independence (or homogeneity), *not* how many people of each race we look at.

Testing independence (I)

- Consider,

Race	Had cold	Didn't have cold	Total
Whites	801	104	905
Non-whites	83	12	95
Total	884	116	1000

  and assume it was obtained fixing only $N = 1000$.
- The hypothesis of interest is $H_0 : p_{ij} = p_{i.} \times p_{.j}$
- $\hat{p}_{11} = 0.884 \times 0.905$, and $E_{11} = 1000 \times 0.884 \times 0.905$. Similarly for the rest.

Testing independence (II)

Race	Had cold	Didn't have cold	Total
Whites	801	104	905
Non-whites	83	12	95
Total	884	116	1000

- Apparently, we estimate 4 parameters p_{ij} for the 4 cells.
- Conditionally on the margins, only two parameters are free, and need to be counted.

Testing independence (III)

O_{ij}	E_{ij}	$(O_{ij} - E_{ij})$	$(O_{ij} - E_{ij})^2$	$\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$
801	800.02	-0.98	0.9604	0.00120
83	83.98	0.98	0.9604	0.01144
104	104.98	0.98	0.9604	0.00915
12	11.02	-0.98	0.9604	0.08715
			$Z =$	0.10894

- The expected values are computed as $E_{ij} = Np_{ij} = Np_{i.}p_{.j}$.
- For instance, $800.02 = 1000 \times 0.884 \times 0.905$.
- Degrees of freedom are $k - p - 1 = 4 - 2 - 1 = 1$. So we have to compare 0.10894 with the quantiles of a χ^2_1 distribution.

Testing independence (IV)

- We can easily construct the table:

```
> ColdRace <- matrix(c(801, 83, 104, 12), 2,
    2)
> ColdRace <- as.table(ColdRace)
> colnames(ColdRace) <- c("Cold", "Not-Cold")
> rownames(ColdRace) <- c("Whites", "Non-whites")
> ColdRace
           Cold Not-Cold
Whites      801      104
Non-whites   83       12
```

Testing independence (V)

Function `loglin` fits, among many other things, the independence model:

```
> result <- loglin(ColdRace, margin = list(1,
    2), fit = TRUE)
2 iterations: deviation 0
> result$pearson
[1] 0.1089356
> result$df
[1] 1
```

Testing independence (VI)

- The E_{ij} approach quite well O_{ij}:

```
> result$fit
             Cold Not-Cold
Whites     800.02   104.98
Non-whites  83.98    11.02
```

- We can now test the independence hypothesis:

```
> 1 - pchisq(result$pearson, df = result$df)
[1] 0.7413592
```

- The tail is 0.7414; there is no reason to reject the independence hypothesis.

Testing homogeneity (I)

- Consider again,

Observed counts ($= O_i$)

Race	Had cold	Didn't have cold	Total
Whites	801	104	905
Non-whites	83	12	95
Total	884	116	1000

but this time assuming we have fixed the row marginal.

- We are testing the hypothesis $H_0 : p_{1j} = p_{2j}$ for all j.
- Under H_0, $\hat{p}_{.j} = n_{.j}/n_{..}$ is a sensible estimate of $p_{.j}$, common to all i.

Testing homogeneity (II)

- The results are exactly the same, only they are arrived at in a different manner.

Expected counts ($= E_i$)

Race	Had cold	Didn't have cold	Total
Whites	800.02	104.98	905
Non-whites	83.98	11.02	95
Total	884	116	1000

- The E_{1j} in the first row are computed as $905 \times \hat{p}_{.j}$
- The E_{2j} in the second row are computed as $95 \times \hat{p}_{.j}$

Testing homogeneity (III)

- $$Z_1 = \sum_{j=1}^{2} \frac{(O_{1j} - E_{1j})^2}{E_{1j}}$$

  for the cells in the first row would be distributed as $\chi^2_{k-1} = \chi^2_1$ if no parameters were estimated and the $p_{.j}$ used were the correct p_{1j}.
- Likewise,
  $$Z_2 = \sum_{j=1}^{2} \frac{(O_{2j} - E_{2j})^2}{E_{2j}}$$
  woud be χ^2_1.
- $Z = Z_1 + Z_2$ would be distributed as a χ^2_2, but we have to subtract **1** parameter $p_{.1}$ estimated (why not also $p_{.2}$?).
- **The same** statistic Z follows **the same** distribution under H_0 than in the case of independence.

General rule

- When testing either independence or homogeneity in an $r \times s$ contingency table, in both cases we form
  $$Z = \sum_{i,j}^{2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$
- The resulting value of Z is (under the null hypothesis of independence or homogeneity) distributed as:
  $$\chi^2_{(r-1)(s-1)}$$
- H_0 should be rejected if Z falls in the α right tail of said distribution (alternatively: if the probability to the right of Z in a $\chi^2_{(r-1)(s-1)}$ is "small").

Fisher's exact test (I)

- Consider again our table,

Race	Had cold	Didn't have cold	Total
Whites	n_{11}	n_{12}	$n_{1.}$
Non-whites	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	$N = n_{..}$

- For given $p_{11}, p_{21}, p_{12}, p_{22}$ its probability would be:
  $$\frac{N!}{n_{11}!n_{12}!n_{21}!n_{22}!} p_{11}^{n_{11}} p_{21}^{n_{21}} p_{12}^{n_{12}} p_{22}^{n_{22}}$$

Fisher's exact test (II)

- The probabilities that N is distributed as it is in the row and column margins are respectively:
  $$\frac{N!}{n_{1.}!n_{2.}!} p_{1.}^{n_{1.}} p_{2.}^{n_{2.}} \qquad \frac{N!}{n_{.1}!n_{.1}!} p_{.1}^{n_{.1}} p_{.2}^{n_{.2}}$$
- Conditional on the margins, the probability of a given table is:
  $$\frac{\left(\frac{N!}{n_{11}!n_{12}!n_{21}!n_{22}!} p_{11}^{n_{11}} p_{21}^{n_{21}} p_{12}^{n_{12}} p_{22}^{n_{22}}\right)}{\left(\frac{N!}{n_{1.}!n_{2.}!} p_{1.}^{n_{1.}} p_{2.}^{n_{2.}}\right)\left(\frac{N!}{n_{.1}!n_{.2}!} p_{.1}^{n_{.1}} p_{.2}^{n_{.2}}\right)}$$
- Under the null hypothesis $p_{ij} = p_{i.}p_{.j}$ all nuisance parameters cancel!

Fisher's exact test (III)

- All we are left with for the probability of a given table is:

$$\frac{\left(\frac{N!}{n_{11}!n_{12}!n_{21}!n_{22}!}\right)}{\left(\frac{N!}{n_{1.}!n_{2.}!}\right)\left(\frac{N!}{n_{.1}!n_{.2}!}\right)}$$

- The denominator is always the same.
- Can compute the probability of each table under the null $H_0 : p_{ij} = p_{i.}p_{.j}$ and check whether what we have observed is very unlikely.
- Unfeasible for large tables.

Fisher's exact test (IV)

- Function to do it in R. Useful for small tables; no approximations. Will fail for large tables.

```
> fisher.test(ColdRace)
        Fisher's Exact Test for Count Data

data:  ColdRace
p-value = 0.7363
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.5345751 2.1400287
sample estimates:
odds ratio
  1.113439
```

Introduction

- Normal distribution is a useful model in many situations.
- Why? Central Limit Theorem.
- Even when the the distribution of a random variable is not normal, normal theory based tests are surprisingly adequate.
- By "adequate" is meant that significance levels (α) and power ($1-\beta$) are close to theoretical values.

$H_0 : m = m_0$ with $X \sim N(m, \sigma^2)$ and σ^2 known (I)

- We have $\overline{X} \sim N(m_0, \sigma^2/n)$ and therefore:

$$T = \frac{\overline{X} - m_0}{\sigma/\sqrt{n}} \sim N(0,1)$$

- T can be computed because σ^2 is known.
- Hence,

$$\text{Prob}\left\{-z_{\alpha/2} \leq T \leq z_{\alpha/2}\right\} = 1 - \alpha$$

- We would reject H_0 at the significance level α if $|T| > |z_{\alpha/2}|$.

$H_0 : m = m_0$ with $X \sim N(m, \sigma^2)$ and σ^2 known (II)

- If we expect departures from H_0 to be of the form $m > m_0$ or $m < m_0$ we would adjust the critical region accordingly:

$$m > m_0 \implies \text{Reject if } T > z_\alpha$$
$$m < m_0 \implies \text{Reject if } T < -z_\alpha$$

- Makes sense when looking at the test statistic $T = \dfrac{\overline{X} - m_0}{\sigma/\sqrt{n}}$; would also be the answer given by the Neyman and Pearson theorem for a simple alternative.
- "Reject if $|T| > |z_{\alpha/2}|$" is just a compromise when no clear alternative.

A digression: confidence intervals

- When testing H_0 with no given alternative, the "unlikely" region is the critical region.
- The "likely" region is the confidence interval.
- This does **not** extend to tests with a prescribed alternative H_a.
- When we have a H_a, the critical region may be one-sided, not the complement of the confidence interval.

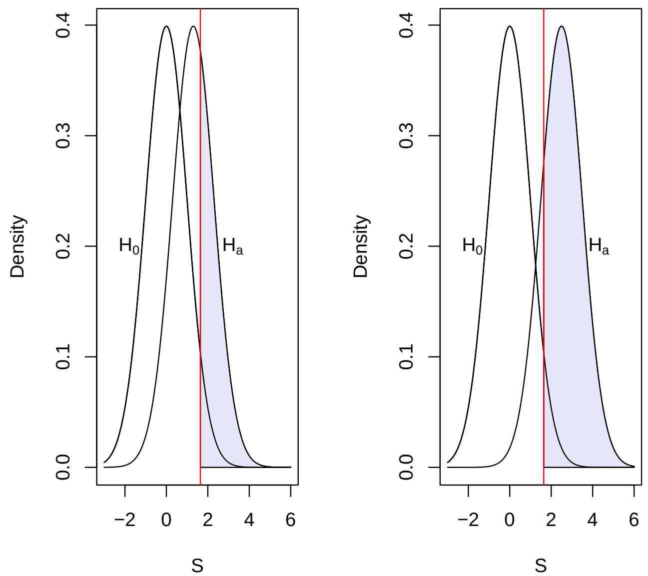
$H_0 : m = m_0$ with $X \sim N(m, \sigma^2)$ and σ^2 known (III)

- What is the payoff of a larger sample size n?
- The test statistic always is $N(0, 1)$ distributed under the null H_0.
- However, under an alternative $m \neq m_0$,

$$T = \frac{\sqrt{n}(\overline{X} - m_0)}{\sigma}$$

has mean $\sqrt{n}(m - m_0)/\sigma$.
- For given m, the greater n, the farther away from 0 is the mean of the test statistic.

$H_0 : m = m_0$ with $X \sim N(m, \sigma^2)$ and σ^2 known (IV)



$H_0 : m = m_0$ with $X \sim N(m, \sigma^2)$ and σ^2 unknown (I)

- Now, $T = \dfrac{\overline{X} - m_0}{\sigma/\sqrt{n}}$ cannot be computed, for σ^2 is not known.
- Replacing σ^2 by its estimate $s^2 = n^{-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$ gives an estimator whose distribution under H_0 is no longer $N(0,1)$.
- Key fact:
$$\frac{nS^2}{\sigma^2} \sim \chi^2_{n-1}$$
and is **independent** of $\overline{X}$.
- This paves the way to eliminating the nuisance parameter σ^2 by *studentization*.

$H_0 : m = m_0$ with $X \sim N(m, \sigma^2)$ and σ^2 unknown (II)

- The ratio,
$$T = \frac{\frac{\sqrt{n}(\overline{X}-m_0)}{\sigma}}{\sqrt{\frac{nS^2/\sigma^2}{n-1}}} = \frac{(\overline{X} - m_0)}{S}\sqrt{n-1} \sim \frac{N(0,1)}{\sqrt{\frac{\chi^2_{n-1}}{n-1}}}$$
when $H_0 : m = m_0$ is true.
- Therefore we can compare the values of the test statistic T to a t_{n-1} (Student's t with $n-1$ degrees of freedom).
- Decision rule: "Reject H_0 if $|T| > t_{\alpha/2;n-1}$."
- Again, we take critical regions of full α size to the right or to the left, if alternative is one-sided.

Example: $H_0 : m_0 = 2$, $\sigma^2 = 1$ known

- Let the sample be:

```
> x <- c(2.2, 3.4, 2.9, 3, 1.6, 3, 3.1, 3.6,
     1.9)
> length(x)
[1] 9
> T <- sqrt(9) * (mean(x) - 2)/1
> T
[1] 2.233333
> qnorm(0.975)
[1] 1.959964
```

- In this case, we would reject.

Example: $H_0 : m_0 = 2$, $\sigma^2 = 1$ unknown

- Now, we would compute

```
> T <- sqrt(9 - 1) * (mean(x) - 2)/sqrt(8 *
     var(x)/9)
> T
[1] 3.256689
> qt(0.975, df = 8)
[1] 2.306004
> var(x)
[1] 0.4702778
> sum((x - mean(x))^2)/9
[1] 0.4180247
> (8 * var(x)/9)
[1] 0.4180247
```

$H_0 : \sigma^2 = \sigma_0^2$ with $X \sim N(m, \sigma^2)$

- Under the null hypothesis,

$$T = \frac{nS^2}{\sigma_0^2} \sim \chi^2_{n-1}$$

- Therefore,

$$\mathrm{Prob}\left\{\chi^2_{n-1;1-\alpha/2} \le \frac{nS^2}{\sigma_0^2} \le \chi^2_{n-1;\alpha/2}\right\} = 1 - \alpha$$

- Critical region $[0, \chi^2_{n-1;1-\alpha/2}] \cup [\chi^2_{n-1;\alpha/2}, \infty)$, unless we have an alternative $H_a : \sigma^2 < \sigma_0^2$ or $H_a : \sigma^2 > \sigma_0^2$
- In the first case the critical region would be $[0, \chi^2_{n-1;1-\alpha}]$, in the second $[\chi^2_{n-1;\alpha}, \infty)$

$H_0 : m_1 - m_2 = m_1^* - m_2^*$ with X, Y normal, variances known (I)

- The commonest test by far is that of $H_0 : m_1 - m_2 = 0$, but we present the test generally.
- We have,

$$\overline{X} - \overline{Y} \sim N\left(m_1 - m_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

- Hence, under H_0,

$$\frac{\overline{X} - \overline{Y} - (m_1^* - m_2^*)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n}}} \sim N(0, 1)$$

$H_0 : m_1 - m_2 = m_1^* - m_2^*$ with X, Y normal, variances known (II)

- Therefore, under H_0,

$$\mathrm{Prob}\left\{-z_{\alpha/2} \le \frac{\overline{X} - \overline{Y} - (m_1^* - m_2^*)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \le z_{\alpha/2}\right\} = 1 - \alpha$$

- The critical region for the test statistic is made of the two $\alpha/2$ tails, unless we have reason to expect the deviance to be one-sided.

$H_0 : m_1 - m_2 = m_1^* - m_2^*$ with X, Y normal, variances $\sigma_1^2 = \sigma_2^2$ unknown (I)

- We have,

$$\overline{X} - \overline{Y} \sim N\left(m_1 - m_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

$$\frac{n_1 S_1^2}{\sigma_1^2} + \frac{n_2 S_2^2}{\sigma_2^2} \sim \chi^2_{n_1+n_2-2}$$

- Using the crucial assumption that $\sigma_1^2 = \sigma_2^2 = \sigma^2$ we can construct a test statistic which does not depend on σ^2 .

$H_0 : m_1 - m_2 = m_1^* - m_2^*$ with X, Y normal, variances $\sigma_1^2 = \sigma_2^2$ unknown (II)

- Using $\sigma_1^2 = \sigma_2^2 = \sigma^2$

$$\frac{\frac{\overline{X}-\overline{Y}-(m_1-m_2)}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}}{\frac{1}{\sigma}\sqrt{\frac{n_1S_1^2+n_2S_2^2}{n_1+n_2-2}}} \sim t_{n_1+n_2-2}$$

- Cancelling the nuisance parameter σ we end up with:

$$\frac{\overline{X}-\overline{Y}-(m_1-m_2)}{\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\sqrt{\frac{n_1S_1^2+n_2S_2^2}{n_1+n_2-2}}} \sim t_{n_1+n_2-2}$$

- Assumption $\sigma_1^2 = \sigma_2^2$ crucial, otherwise an open question (so-called Behrens-Fisher problem).

$H_0 : \sigma_1^2/\sigma_2^2 = \sigma_{1*}^2/\sigma_{2*}^2$ with X, Y normal (I)

- With respective sample sizes n_1 and n_2, we have:

$$\frac{n_1S_1^2}{\sigma_1^2} \sim \chi^2_{n_1-1} \qquad \frac{n_2S_2^2}{\sigma_2^2} \sim \chi^2_{n_2-1}$$

- Clearly both statistics are independent, so

$$\frac{n_1S_1^2\sigma_2^2(n_2-1)}{n_2S_2^2\sigma_1^2(n_1-1)} \sim \mathcal{F}_{n_1-1,n_2-1}$$

- It the hypothesis H_0 is true, replacing σ_1^2, σ_2^2 by their hypothetical values would give a test statistic with the distribution shown.

General ideas

- Tests for a mean or the difference of means are remarkably robust to deviations from normality; however, to play safe we might use tests to be described next.
- Tests for the difference of means are quite sensitive to different variances: the requirement $\sigma_1^2 = \sigma_2^2$ cannot be dispensed with.

Permutation tests (I)

- Easy alternative when distribution cannot be assumed and we can use a computer.
- Want to test $x_1, \ldots, x_{n_1}$ and $y_1, \ldots, y_{n_2}$ are indeed samples form the same population, the alternative being that the means are different.
- Our test statistic is $\overline{x} - \overline{y}$. Need something to compare to.
- If we arrange the observations as:

$$x_1, \ldots, x_{n_1}, y_1, \ldots, y_{n_2}$$

$\overline{x} - \overline{y}$ is just the difference of the averages of the first n_1 and subsequent n_2 observations.

Permutation tests (II)

- If observations come indeed from the same population, the difference between the n_1 and n_2 observations in each group should be of similar magnitude than that among any other set of n_1 and n_2 observations.
- Idea: sample repeatedly the whose set of observations in random subsets of n_1 and n_2, and compute each time $(\overline{x} - \overline{y})_j$ $(j = 1, \ldots, N)$.
- Compare the observed $\overline{x} - \overline{y}$ to $(\overline{x} - \overline{y})_j$ $(j = 1, \ldots, N)$ and reject H_0 if it is in an extreme position.
- Sampling is usually done by permuting the original sample, hence the name.

Testing $H_0 : m = m_0$ with no normality (I)

- For "large" n (=sample size), use normal theory tests. "Large" is $n \geq 30$ (if σ^2 is known) and $n \geq 100$ (if it is not).
- For smaller n, remember Tchebycheff inequality:

$$\mathrm{Prob}\left\{|X - m| < k\sigma\right\} \geq 1 - \frac{1}{k^2}$$

- For the particular case of $\overline{X}$ we have:

$$\mathrm{Prob}\left\{|\overline{X} - m| < \frac{k\sigma}{\sqrt{n}}\right\} \geq 1 - \frac{1}{k^2}$$

Testing $H_0 : m = m_0$ with no normality (II)

- Therefore, replacing k by $1/\sqrt{\alpha}$ we have:

$$\mathrm{Prob}\left\{|\overline{X} - m| < \frac{\sigma}{\sqrt{n\alpha}}\right\} \geq 1 - \alpha$$

- This gives as a basis for a confidence interval for m and a test: "Reject H_0 at the α significance level if $|\overline{X} - m_0| > \sigma/\sqrt{n\alpha}$."
- If σ^2 is unknown, replace it by its estimate s^2 to have an approximate test.
- This distribution-free method gives tests less powerful (and confidence intervals wider) than the normal theory tests.

Testing $H_0 : m = m_0$ with no normality (III)

- If the sample size n is large enough, the statistics

$$\frac{\overline{X} - m}{\sigma/\sqrt{n}} \qquad \frac{\overline{X} - m}{s/\sqrt{n}}$$

have approximate $N(0,1)$ distributions, even if X is not normal.
- "Large" means $n \geq 30$ if σ is known and $n \geq 100$ if we are forced to use s instead.
- The approximation is usually quite good, and can be checked by simulation (e.g., repeatedly generate samples of size n from, say, the uniform distribution, and plot the histogram of values of the test statistic; except in pathological cases, it will approach a normal bell curve shape).

The case of a proportion (I)

- One case of particular interest is that of a proportion. Variable X the value 0 or 1 ("yes" or "no", or similar dichotomous values coded to 1/0).
- We are interested in the probability of 1, p.
- Clearly $\overline{X} = n^{-1}(X_1 + \ldots + X_n)$ is an unbiased estimate of p.
- How to test hypothesis on p or estimate it by interval? We know that for large n approximately,

$$\frac{\overline{X} - p}{s/\sqrt{n}} \approx N(0,1)$$

- We can estimate s^2 by $\hat{p}(1-\hat{p})$ or (conservatively) by 0.25.
- However we estimate p, approximately, for large n,

$$\frac{\overline{X} - m}{s/\sqrt{n}} \approx N(0,1)$$

How would we construct a confidence interval for p

$(\overline{X} \pm z_{\alpha/2} s/\sqrt{n})$

The case of a proportion (II)

Example:

- In a sample of 500 parts from a very large batch, 33 are found to be defective. Would the hypothesis $H_0 : p = 0.04$ be rejected against an alternative $H_a : p > 0.04$? ($\alpha = 0.05$).
- The estimate of p would be $33/500 = 0.0666$ and $s^2 = pq = 0.04 \times 0.96 = 0.0384$. Under H_0,

$$\frac{(\overline{X} - 0.04)}{\sqrt{0.0384}/\sqrt{500}} \approx N(0,1);$$

the critical region would be to the right.
- Replacing $\overline{X}$ by 33/500 we get a value for the test statistic of 3.04, well inside a critical region of size $\alpha = 0.01$. So we would reject H_0 at said level of significance.

The case of a proportion (III)

Example (continued):

- If we were asked to estimate by interval the true p with confidence $1 - \alpha = 0.99$, we could use:

$$\frac{(\overline{X} - p)}{\sqrt{\frac{0.0667 \times 0.9333}{500}}} \approx N(0,1)$$

- Then,

$$\text{Prob}\left\{\overline{X} - 2.5758\sqrt{\frac{0.06222}{500}} \leq p \leq \overline{X} + 2.5758\sqrt{\frac{0.06222}{500}}\right\} \approx 0.99$$

- The confidence interval would thus be (0.0666 ± 0.0287)
- Replacing s^2 by the upper bound of $p(1-p) = 0.25$ would be *very* conservative here.

Testing differences of means

- We state without proof the following approximate results:

$$\frac{\overline{X} - \overline{Y} - (m_1 - m_2)}{\sqrt{\frac{\sigma^2{}_1}{n_1} + \frac{\sigma^2{}_2}{n_2}}} \approx N(0,1) \qquad (n_1 \geq 30, n_2 \geq 30)$$

$$\frac{\overline{X} - \overline{Y} - (m_1 - m_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \approx N(0,1) \qquad (n_1 \geq 100, n_2 \geq 100)$$

- Those approximate distributions can be used in the construction of test statistics or confidence intervals.

Testing differences of proportions

- The results in the previous slide can be specialized to the case of two proportions. In that case,

$$\overline{X} = \frac{Z_1}{n_1} \qquad m_1 = p_1$$
$$\overline{Y} = \frac{Z_2}{n_2} \qquad m_2 = p_2$$
$$\frac{\frac{Z_1}{n_1} - \frac{Z_2}{n_2} - (p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} \approx N(0,1)$$

- Again, sample sizes should be large.

How would we construct a confidence interval for $(p_1 - p_2)$?

$$\left(\frac{Z_1}{n_1} - \frac{Z_2}{n_2}\right) \pm z_{\alpha/2}\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

The OC ("operating characteristic") curve (I)

- The performance of a test of H_0 against a set of alternatives usually described by the OC curve: it gives the probability of non-rejection of H_0 for both the null and a range of alternatives.
- Common in specification of industrial quality sampling protocols.
- The conflicting interests ob buyer and seller are specified in two points, through which the curve is forced to pass.

The OC ("operating characteristic") curve (II)