Last time :
- Sample size ≤ 10, can use sample range
- usually, use s^2 to estimate σ^2
- chi-square distribution
- confidence interval for σ^2: $\left[\frac{(n-1)s^2}{\chi^2_{\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}\right]$
- hypothesis testing $\chi^2 = \frac{(n-1)s^2}{\sigma_0}$

If we have a two tailed test, we can use a confidence interval to perform the test.

- e.g. Statistics Canada claims that the standard deviation in the heights of Canadian women is 5cm. We randomly select 11 women, and put a sample standard deviation at 6cm. Test the claim at a .05 level of significance.

$H_0 : \sigma^2 = 5^2$
$H_1 : \sigma^2 \neq 5^2$

The $100(1-\alpha)\% = 95\%$ confidence interval for σ^2 is $\left[\frac{10(6)^2}{\chi^2_{.025}}, \frac{10(6)^2}{\chi^2_{.975}}\right]$

$$\Rightarrow \left[\frac{10(6)^2}{20.483}, \frac{10(6)^2}{3.247}\right] = (17.58, 110.87)$$

As 5^2 lies in the 95% confidence interval, we cannot reject the claim at a .05 level of confidence.

- Suppose we have two normal populations, and we want to compare their variances, $Pop_1 = \sigma_1^2$
$Pop_2 = \sigma_2^2$

If we make the assumption that the variances are the ... and we take a random sample of size n_1 from Pop_1, n_2 from Pop_2, and let $F = \frac{S_1^2}{S_2^2}$

Then F has distribution with $\nu_1 = n_1 - 1$, $\nu_2 = n_2 - 1$

We have $F_{.01}$ and $F_{.05}$ values.

$Pr(F > F_\alpha) = \alpha$

$\left.\begin{array}{l} H_0 : \sigma_1^2 \leq \sigma_2^2 \\ H_1 : \sigma_1^2 > \sigma_2^2 \end{array}\right\}$ $F = \frac{S_1^2}{S_2^2}$, reject that if $F > F_\alpha$

$\left.\begin{array}{l} H_0 : \sigma_1^2 \geq \sigma_2^2 \\ H_1 : \sigma_1^2 < \sigma_2^2 \end{array}\right\}$ swap pops 1 + 2

$H_0 : \sigma_1^2 = \sigma_2^2$ } IF necessary, swap so that $S_1^2 \geq S_2^2$
$H_1 : \sigma_1^2 \neq \sigma_2^2$ } Repeat that if $F > F_{\alpha/2}$

- e.g. Krusty claims that the standard deviation in the number of marshmallows in boxes of Lucky charms is at least as great as that in Krusty-O's. We randomly select 10 boxes of Lucky Charms and 8 boxes of Krusty O's

For the lucky charms, we get a sample standard deviation of 10, for the Krusty O's, 15.

Test the claim at a .05 level of significance.

Pop₁ = Krusty-O's   Pop₂ = Lucky Charms

$H_0 : \sigma_1^2 \leq \sigma_2^2$ $F = \frac{S_1^2}{S_2^2}$ and reject that if $F > F_{.05}$

$H_1 : \sigma_1^2 > \sigma_2^2$

$$F = \frac{15^2}{10^2} = 2.25$$

As $\nu_1 = 7$, $\nu_2 = 9$, $F_{.05} = 3.29$

As $F \not> F_{.05}$, we cannot reject the claim at a .05 level of significance.

Chapter 11 – Repression Analysis

Let x and y be random variables

y will depend upon x plus a randomness factor

$y = f(x) + \varepsilon$, when f is a function and ε is a random variable with non-0. We call $f(x)$ the <u>regression curve</u>. Finding it is called <u>repression analysis</u>.

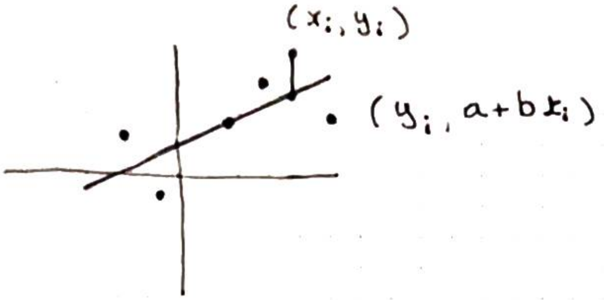
We can get an idea by taking a random sample $(x_1, y_1), \ldots, (x_n, y_n)$, and drawing a scatterplot.

In <u>linear regression</u>, we have $y = \alpha + \beta x + \varepsilon$, $\alpha, \beta \in \mathbb{R}$

We will find the <u>line of best fit</u> for our data,

$\tilde{y} = a + bx$

We will use the <u>method of least squares</u>



We want to minimize the sum of the squares of the vertical distance from the points to the line

$$\sum_{i=1}^{n} (y_i - (a+bx_i))^2$$

$$a : \sum_{i=1}^{n} 2(y_i - (a+bx_i))(-1) = 0$$

$$\sum_{i=1}^{n} a + \sum_{i=1}^{n} bx_i = \sum_{i=1}^{n} y_i$$

$$na + \left(\sum_{i=1}^{n} x_i\right) b = \sum_{i=1}^{n} y_i \quad (1)$$

$$b : \sum_{i=1}^{n} 2(y_i - (a+bx_i))(-x_i) = 0$$

$$\sum_{i=1}^{n} ax_i + \sum_{i=1}^{n} bx_i^2 = \sum_{i=1}^{n} x_i y_i$$

$$\left(\sum_{i=1}^{n} x_i\right) a + \left(\sum_{i=1}^{n} x_i^2\right) b = \sum_{i=1}^{n} x_i y_i \quad (2)$$

Solve (1), (2) for a, b

- e.g. Find the line of best fit for:

x_i	1	2	3
y_i	7	3	1

$\rightarrow na + \sum x_i b = \sum y_i$

(1) $3a + 6b = 11$

$\rightarrow \sum x_i a + \sum x_i^2 b = \sum x_i y_i$

(2) $6a + 14b = 16$

(2) $- 2$(1) $\Rightarrow 2b = -6$, $b = -3$

Subbing into (1): $3a + 6(-3) = 11$, $a = \frac{29}{3}$

$$\tilde{y} = \frac{29}{3} - 3x$$

- in the above example, estimating when $x = 2.5$

$$\tilde{y} = \frac{29}{3} - 3(2.5)$$

- Find the line of best fit for:

x_i	1	2	3	4
y_i	2	4	5	7

→ $na + \sum x_i b = \sum y_i$

① $4a + 10b = 18$

→ $\sum x_i a + \sum x_i^2 b = \sum x_i y_i$

② $10a + 30b = 53$

② − 3① : $-2a = -1$

$a = 1/2$

Subbing into ① : $4(1/2) + 10b = 18$

$b = 1.6$

$\tilde{y} = (0.5) + (1.6)x$

April 6/18
Applied Anal.
TUTORIAL

- e.g. we have a normal population with $\sigma = 10$
Suppose we wish to test.

$H_0 : \mu = 100$ we will take a random sample of
$H_1 : \mu > 100$ size 25 and reject H_0 if $\bar{x} > 104$
Find α.

We assume $\mu = 100$

$$Pr(\bar{x} > 104) = 1 - Pr(\bar{x} \le 104)$$
$$= 1 - F\left(\frac{104-100}{10/\sqrt{25}}\right) = 1 - F(2) = 1 - 0.9772$$
$$= 0.0228$$

- e.g. binomial : bernouli trials
(Flipping coin, rolling die) } replacement!

- e.g. At least 4 hearts, and 1 seven?

4 hearts, including 7
$\binom{12}{3} 36$

4 hearts, not including 7
$\binom{12}{4} 3$

5 hearts
$\binom{12}{4}$

$$\Rightarrow \frac{\binom{12}{3} 36 + \binom{12}{4} 3 + \binom{12}{4}}{\binom{52}{5}}$$

- e.g. Two kings and one club?

King of clubs, another king
$3 \binom{36}{4}$

no king of clubs
$\binom{3}{2}(12)\binom{36}{2}$

$$\Rightarrow \frac{3\binom{36}{4} + \binom{3}{2} 12 \binom{36}{2}}{\binom{52}{5}}$$

- e.g. A bag contains 40 red marbles and 60 blue marbles. We reach into the bag and pull out 10 marbles. Find the prob that we get 3 red marbles.

$$\frac{\binom{80}{3}\binom{60}{7}}{\binom{100}{10}}$$

$N = 100$
$n = 10$
$G = 40$
$x = 3$

w/ replacement: $\binom{10}{3}\left(\frac{80}{100}\right)^3\left(\frac{60}{100}\right)^3$

- e.g. Roll a balanced die 20 times, count the threes
  (i) Find the prob. of at most 4 threes
  (ii) Find the prob. of at least 4 threes

  (i) $B(4; 20, 1/6)$

  (ii) $1 - B(3, -20, 1/6)$

- e.g. Krusty claims that the average box of Krusty-O's contains at least 60 marshmallows. We randomly select 10 boxes and put a sample mean of 56 and a sample standard deviation at 10.
  Test the claim at a .05 level of significance

$H_0 : \mu \geq 60$ As σ is unknown, $n < 30$

$H_1 : \mu < 60$ $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$, reject H_0 if $t < t_{.05}$

$t = \dfrac{56-60}{10/\sqrt{16}} = -1.6$, As $\nu = 15$ $t_{.05} = 1.753$

As $t \nless -t_{.05}$ we cannot reject the claim.

- e.g. Krusty claims that the standard deviation in the number of jagged metal Krusty O's per box is 5. We randomly select 41 boxes and get a standard deviation of 3.
  Test the claim at a .05 level of significance

$H_0 : \sigma^2 = 5^2$ $\chi^2 = \dfrac{(n-1)s^2}{\sigma_0^2}$, reject if $\chi^2 > \chi^2_{.025}$

$H_1 : \sigma^2 \neq 5^2$

$\chi^2 = \dfrac{40(3)^2}{5^2} = 14.4$

As $\nu = 4$, $\chi^2_{.025} = 59.342$

$\chi^2_{.975} = 24.433$

As $\chi^2 = \chi^2_{.975}$, we reject H_0

- e.g. Find the line of best fit:

x_i	1	2	3	4
y_i	6	5	3	1

$na + \sum x_i b = \sum y_i$

$4a + 10b = 15$ ①

$\sum x_i a + \sum x_i^2 b = \sum x_i y_i$

$10a + (1+4+9+16)b = (6+10+9+4)$

$10a + 30b = 29$ ②

solving ① with ②

② - 3① $= -2a = -16$

$a = 8$

then, $b = -1.4$

$\tilde{y} = 8 - 1.4x$

APRIL 6/18
APPLIED ANAL.

- Last time
  - Confidence intervals for two-handed test
  - F-test for two variables
  - regression analysis $y = f(x) + \varepsilon$
  - $y = \alpha + \beta x + \varepsilon$
  - line of best fit: $\hat{y} = a + bx$

$$na + \sum x_i b = \sum y_i$$
$$\sum x_i a + \sum x_i^2 b = \sum x_i y_i$$

If our data is $(x_1, y_1), \dots, (x_n, y_n)$

$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$, $S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$, $S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$

$b = S_{xy}/S_{xx} \Rightarrow a = \bar{y} - b\bar{x}$

- e.g.

x_i	1	2	3
y_i	8	3	1

$\bar{x} = \frac{1+2+3}{3} = 2$, $S_{xx} = ((1-2)^2 + (2-2)^2 + (3-2)^2) = 2$

$\bar{y} = \frac{8+3+1}{3} = 4$, $S_{yy} = ((8-4)^2 + (3-4)^2 + (1-4)^2) = 26$

$S_{xy} = (1-2)(8-4) + (2-2)(3-4) + (3-2)(1-4) = -4 + 0 - 3 = -7$

$b = S_{xy}/S_{xx} \Rightarrow (-7)/(2) = -7/2$; $a = 4 - (\frac{-7}{2})(2) \Rightarrow 4 + 7 = 11$

$\hat{y} = 11 - \frac{7}{2}x$

<u>Exponential</u> regression : $y = \alpha\beta^x + \varepsilon$

Our best fit curve will be $\hat{y} = ab^x$

$\log\hat{y} = \log a + x\log b$

Let $c = \log a$, $d = \log b$, we take $\log\hat{y} = c + dx$

We perform linear regression with $(x_i, \log y_i)$

x_i	1	2	3
y_i	1	10	1000
$\log y_i$	0	1	3

$nc + \sum x_i d = \sum \log y_i$

$3c + 6d = 4$ (1)

$\sum x_i c + \sum x_i^2 d = \sum x_i \log y_i$

$6c + 14d = 11$ (2)

(2) - 2(1) : $2d = 3$, $d = 3/2$

then $c = -5/3$

$a = 10^c = 10^{-5/3}$, $b = 10^d = 10^{3/2}$

$\hat{y} = 10^{-5/3}(10^{3/2})^x$

Power repression : $y = \alpha x^{\beta} + \varepsilon$

our best Fit curve : $\hat{y} = a x^{b}$

$\log \hat{y} = \log c + b \log x$. Let $c = \log$

$\log \hat{y} = c + d \log x$ Perform linear regression with $(\log x_i, \log y_i)$

x_i	10	100	1000
y_i	1000	100	1
$\log x_i$	1	2	3
$\log y_i$	3	2	0

$nc + \sum \log x_i d = \sum \log y_i$

$3c + 6d = 5$ (1)

$\sum \log x_i c + \sum \log x_i^2 d = \sum (\log x_i)(\log y_i)$

$6c + 14d = 7$ (2)

(2) - 2(1) : $2d = -3$; $d = -3/2$, $3c + 6(\frac{-3}{2}) = 5$, $c = 14/3$

$a = 10^{c} \Rightarrow 10^{14/3}$; $b = d = -3/2$

$\hat{y} = 10^{14/3} x^{-3/2}$

Reciprocal regression : $y = \frac{1}{\alpha + \beta x} + \varepsilon$

Our best Fit curve is $\hat{y} = \frac{1}{a + bx}$

$1/\hat{y} = a + bx$. Perform linear regression with $(x_i, 1/y_i)$

- e.g.

x_i	1	2	3	4
y_i	1	1/3	1/4	1/5
$1/y_i$	1	3	4	5

$na + \sum x_i b = \sum 1/y_i$

$4a + 10b = 13$ (1)

$\sum x_i a + \sum x_i^2 b = \sum \frac{x_i}{y_i}$

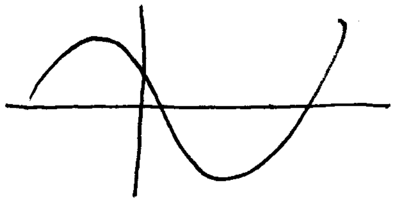
$10a + 30b = 39$ (2)

3(1) - (2) : $2a = 0$, $a = 0$; $b = 39/30 = 13/10$

$\hat{y} = \frac{1}{(13/10) x}$

Polynomial regression: $y = \beta_0 + \beta_1 x + \beta_2 x^2, \dots, \beta_p x^p + \varepsilon$

Our best fit curve is $\hat{y} = S_0 + S_1 x + S_2 x^2 + \dots + S_p x^p$



We will minimize the sum of the sequence of the vertical distance from the points to the curve.

$$\sum_{i=1}^{n} \left(y_i - (b_0 + b_1 x_i + b_2 x_i^2 + \dots + b_p x_i^p)\right)^2$$

Diff art b_0 : $\sum_{i=1}^{n} 2(y_i - (b_0 + b_1 x_i + b_2 x_i^2 + \dots + b_p x_i^p))(-1) = 0$

$$n b_0 + \left(\sum_{i=1}^{n} x_i\right) b_1 + \left(\sum_{i=1}^{n} x_i^2\right) b_2 + \dots + \left(\sum_{i=1}^{n} x_i^p\right) b_p = \sum_{i=1}^{n} y_i \quad (1)$$

Diff art b_1 : $\sum_{i=1}^{n} 2(y_i - (b_0 + b_1 x_i + b_2 x_i^2 + \dots + b_p x_i^p))(-x_i) = 0$

$$\left(\sum_{i=1}^{n} x_i\right) b_0 + \left(\sum_{i=1}^{n} x_i^2\right) b_1 + \left(\sum_{i=1}^{n} x_i^3\right) b_2 + \dots + \left(\sum_{i=1}^{n} x_i^{p+1}\right) b_p = \sum_{i=1}^{n} x_i y_i \quad (2)$$

$$\left(\sum_{i=1}^{n} x_i^2\right) b_0 + \left(\sum_{i=1}^{n} x_i^3\right) b_1 + \left(\sum_{i=1}^{n} x_i^4\right) b_2 + \dots + \left(\sum_{i=1}^{n} x_i^{p+2}\right) b_p = \sum_{i=1}^{n} x_i^2 y_i \quad (3)$$

$\vdots$

$$\left(\sum_{i=1}^{n} x_i^p\right) b_0 + \left(\sum_{i=1}^{n} x_i^{n+1}\right) b_1 + \left(\sum_{i=1}^{n} x_i^{n+2}\right) b_2 + \dots + \left(\sum_{i=1}^{n} x_i^{2p}\right) b_p = \sum_{i=1}^{n} x_i^p y_i \quad (p+1)$$

Solve for $b_0, b_1, \dots, b_p$

-e.g. Find the quadratic of best fit for:

x_i	1	2	3	4
y_i	-2	0	3	10

$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$

$\hat{y} = b_0 + b_1 x + b_2 x^2$

$n b_0 + \sum x_i b_1 + \sum x_i^2 b_2 = \sum y_i$

$4 b_0 + 10 b_1 + 30 b_2 = 11 \quad (1)$

$\sum x_i b_0 + \sum x_i^2 b_1 + \sum x_i^3 b_2 = \sum x_i y_i$

$10 b_0 + 30 b_1 + 100 b_2 = 47 \quad (2)$

$\sum x_i^2 b_0 + \sum x_i^3 b_1 + \sum x_i^4 b_2 = \sum x_i^2 y_i$

$30 b_0 + 100 b_1 + 354 b_2 = 185 \quad (3)$