STATISTICS

Major formulas

Anna Matysiak

Warsaw School of Economics

based on J.T. Mc Clave, P.G. Benson,T. Sincich: Statistics for Business and Economics, 11th Edition, 2010

Descriptive statistics

Mean

General formula for sample mean (individual data):

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

n – number of elements
x_i – single measure for element i

Formulas for sample mean on grouped data:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i n_i}{n}$$

n – number of all elements
n_i – number of elements in i-th interval
x_i – single measure for element i

$$\overline{x} = \frac{\sum_{i=1}^{n} \mathring{x}_i n_i}{n}$$

if data is grouped in intervals and we do not have access to individual data we use centers of the intervals $\mathring{x}_i$

Mode

is a measurement that occurs most frequent in the data

Median

Calculation of the median (procedure):

Arrange n measuremnts from smallest to largest,

1. If n is odd, m is the middle number
2. If n is even m is the mean of middle two numbers

Percentiles

For any set of n measurements (arranged in ascending order), the ***pth* percentile** is the number such that p% of the measurements fall below the p th percentile and (100-p) % fall above it.

Quartiles

The lower quartile Q_L (Q_1) is the 25th percentile of a dataset. **The middle quartile Me** (Q_2)is the **Median**.

The upper quartile Q_U (Q_3) is the 75th percentile of a dataset

Range

is equal to the largest measurement minus the smallest measurement

Formula: R= $x_{\max} - x_{\min}$

Sample variance

General formula for sample variance (individual data):

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$

n – number of elements

x_i – single measure for element i

$\overline{x}$ - sample mean

Formulas for sample mean on grouped data:

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2 \cdot n_i}{n-1}$$

n – number of all elements

n_i – number of elements in i-th interval

x_i – single measure for element i

$$S^2 = \frac{\sum_{i=1}^{n}(\mathring{x}_i - \overline{x})^2 \cdot n_i}{n-1}$$

if data is grouped in intervals and we do not have access to individual data we use centers of the intervals $\mathring{x}_i$

Sample standard deviation

Formula: $$S = \sqrt{S^2}$$

Coefficient of variance

Formula: $$V = \frac{S}{\bar{x}} \cdot 100\%$$

Interquartile range and devation

Variabiliy measures can be also computed on the basis of quartiles.

Interquartile range: $IRQ = Q_3 - Q_1$

Interquartile deviation: $Q = \dfrac{Q_3 - Q_1}{2}$

Skewness measures

Basic skewness measure:

$$A = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{S} \right)^3$$

n – number of elements
x_i – single measure for element i
$\overline{x}$ - sample mean
S – standard deviation

Other skewness measures:

$$A_1 = \frac{\overline{x} - do}{S}$$

do – mode

$$A_2 = \frac{Q_1 - me + Q_3 - me}{Q_3 - Q_1}$$

Q1, me, Q3 - quartiles

Symmetric distribution: A=0
Rightward Skewness: A>0
Leftward Skewness: A<0

Standardisation

- Computing z-scores

$$Z = \frac{x - \bar{x}}{S}$$

Detecting outliers

Two methods:

1.

An outlier is a value which lies beyond the interval:

$Q_L - 1.5(Q_u - Q_L)$; $Q_U + 1.5(Q_u - Q_L)$

2.

An outlier is a value which lies beyond the interval:

$\overline{x} - 3S;$ $\overline{x} + 3S$

Random variable
and probability distributions

Mean and variance

- The **mean** or **expected value**

$$\mu = EX = \sum x \cdot p(x)$$

- **The variance**

$$\sigma^2 = E\left[(x-\mu)^2\right] = \sum (x-\mu)^2 \cdot p(x)$$

- **Standard deviation**

$$\sigma = \sqrt{\sigma^2}$$

Binomial distribution

- n identical trials
- Only 2 possible outcomes on each trial: S- success and F – failure
- The binomial random variable is the number of successes x
- The probability of x successes is computed as:

$$P(X = x) = \binom{n}{x} p^{x} (1 - p)^{n-x}$$

Binomial distribution

- **Mean** $$\mu = np$$
- **Variance** $$\sigma^2 = np(1-p)$$
- **Standard deviation** $$\sigma = \sqrt{np(1-p)}$$

Probability distribution

for a continuous random variable is represented by a **probability density function** (pdf) *f(x)*.

$$f\left(x\right)\geq 0,$$

$$\int_{a}^{b} f\left(x\right)dx = P\left(a < X < b\right) = P\left(a \leq X \leq b\right)$$

$$P(X = a) = \int_{a}^{a} f\left(x\right)dx = 0$$

$$\int_{-\infty}^{+\infty} f\left(x\right)dx = P\left(-\infty < X \leq +\infty\right) = 1$$

Mean and variance

- **The mean** or **expected value**

$$\mu = \int_{-\infty}^{\infty} x f(x)\, dx$$

- **The variance**

$$\sigma^2 = \int_{-\infty}^{\infty} [x-\mu]^2 f(x)\, dx$$

- **The standard deviation**

$$\sigma = \sqrt{\sigma^2}$$

Normal distribution

- characterised by two parameters, μ (mean) and σ (standard deviation),
- its density function has a form: $X \sim N(\mu, \sigma)$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}}, \qquad -\infty < x < +\infty$$

Normal distribution

$\mu \pm 3\sigma$ rule

$$P(|X-\mu| \leq \sigma) = 0.6827$$

$$P(|X-\mu| \leq 2\sigma) = 0.9545$$

$$P(|X-\mu| \leq 3\sigma) = 0.9973$$

Standard normal distribution

- A normal distribution with $\mu = 0$ and $\sigma = 1$

Standardisation

If $X \sim N(\mu, \sigma)$

then:

$$Z = \frac{X - \mu}{\sigma} \sim N(0,1)$$

Sample and exact and asymptotic sampling distributions

Sampling distribution of a mean

Population distribution	Population standard deviation	Sample size	Sample statistic
normal	known	any	$\bar{x} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$
normal	unknown	>=30	$\bar{x} \sim N(\mu, \frac{S}{\sqrt{n}})$
normal	unknown	< 30	$\frac{\bar{x} - \mu}{\frac{S}{\sqrt{n}}} \sim t - Student$
any	known or unknown	>=100	$\bar{x} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$

Sampling distribution of a sum

Population distribution	Population standard deviation	Sample size	Sample statistic
normal	known	any	$\sum X_i \sim N(\mu \cdot n, \sigma \cdot \sqrt{n})$
normal	unknown	>=30	$\sum X_i \sim N(\mu \cdot n, S \cdot \sqrt{n})$
any	known or unknown	>=100	$\sum X_i \sim N(\mu \cdot n, \sigma \cdot \sqrt{n})$

Sampling distribution of a proportion

$$X \sim N(np, \sqrt{np(1-p)})$$

$$\hat{p} = \frac{X}{n}$$

$$\hat{p} \sim N(p, \sqrt{\frac{p(1-p)}{n}})$$

Estimation

Point Estimators

$$\bar{x} = \frac{\sum_i x_i}{n}$$ minimum variance unbiased estimator of μ

$$\hat{p} = \frac{x}{n}$$ minimum variance unbiased estimator of p

$$\hat{S} = \sqrt{\frac{1}{n-1}\sum (x_i - x)^2}$$ minimum variance unbiased estimator of σ

$$\hat{\tilde{S}} = \sqrt{\frac{1}{n}\sum (x_i - x)^2}$$ biased estimator in smaller samples
(asymptotically unbiased)

Confidence interval for population mean

- If a sample was drawn from a normally distributed population with known standard deviation

$$P(\bar{x} - z_{\alpha/2} \cdot \sigma / \sqrt{n} < \mu < \bar{x} + z_{\alpha/2} \cdot \sigma / \sqrt{n}) = 1 - \alpha$$

- If a sample was drawn from a normally distributed population with unknown standard deviation
  - Large sample

$$P(\bar{x} - z_{\alpha/2} \cdot S / \sqrt{n} < \mu < \bar{x} + z_{\alpha/2} \cdot S / \sqrt{n}) = 1 - \alpha$$

  - Small sample

$$P(\bar{x} - t_{\alpha/2;n-1} \cdot S / \sqrt{n} < \mu < \bar{x} + t_{\alpha/2;n-1} \cdot S / \sqrt{n}) = 1 - \alpha$$

- If a sample was drawn from a population with unknown distribution

$$P(\bar{x} - z_{\alpha/2} \cdot S / \sqrt{n} < \mu < \bar{x} + z_{\alpha/2} \cdot S / \sqrt{n}) = 1 - \alpha$$

Confidence interval for population proportion

$$P\left(\hat{p} - z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) = 1 - \alpha$$

Determining the sample size

Mean:

$$n = \frac{(z_{\alpha/2})^2 \sigma^2}{SE^2}$$

Proportion:

$$n = \frac{(z_{\alpha/2})^2 p(1-p)}{SE^2}$$

where SE – sampling error

Statistical testing
Parametric significance tests.

Inference about
a population parameter

One-tailed vs. two tailed test

Lower-tail test	Upper-tail test	Two-tailed test
$H0: \theta = \theta_0$ $Ha: \theta < \theta_0$	$H0: \theta = \theta_0$ $Ha: \theta > \theta_0$	$H0: \theta = \theta_0$ $Ha: \theta \neq \theta_0$
P-value = P(Z<-z) P-value = P(T<-t)	P-value = P(Z>z) P-value = P(T>t)	P-value = P(Z<-z)+P(Z>z) P-value = P(T<-t)+P(T>t)

Hypothesis about population mean

Situation	Test statistic
sample drawn from a normally distributed population with known standard deviation	$z = \dfrac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$
sample (n>=30) drawn from a normally distributed population with unknown standard deviation	$z = \dfrac{\bar{x} - \mu_0}{S / \sqrt{n}}$
sample (n<30) drawn from a normally distributed population with unknown standard deviation	$t = \dfrac{\bar{x} - \mu_0}{S / \sqrt{n}} \quad df = n - 1$
sample drawn from a population with unknown distribution	$z = \dfrac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$

Hypothesis about population proportion

Test statistic: $$z = \frac{\hat{p} - p_0}{\sigma_p}$$ $$\sigma_p = \sqrt{\frac{p_0(1 - p_0)}{n}}$$

Inference about a difference between parameters from two populations

One-tailed vs. two tailed test

Lower-tail test	Upper-tail test	Two-tailed test
$H0:\theta_1=\theta_2$ $Ha:\theta_1<\theta_2$	$H0:\theta_1=\theta_2$ $Ha:\theta_1>\theta_2$	$H0:\theta_1=\theta_2$ $Ha:\theta_1\neq\theta_2$
P-value = P(Z<-z) P-value = P(T<-t)	P-value = P(Z>z) P-value = P(T>t)	P-value = P(Z<-z)+P(Z>z) P-value = P(T<-t)+P(T>tt)

Hypothesis about two population means

Situation	Test statistic
samples drawn from two normally distributed populations with known standard deviations	$z = \dfrac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$
samples (n>=30) drawn from two normally distributed populations with unknown standard deviations	$z = \dfrac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}}$
samples (n<30) drawn from two normally distributed populations with unknown standard deviations	$t = \dfrac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{S_p^2\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$ $S_p^2 = \dfrac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$ $df = n_1 + n_2 - 2$
samples drawn from two populations with unknown distributions	$z = \dfrac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$

Hypothesis about population proportion

Test statistic:

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - D_0}{\sqrt{\tilde{p}(1-\tilde{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where:

$$\tilde{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

Non-parametric tests

Chi-square test for assessing normality

H0: $F(x) = F0(x)$
Ha: $F(x) \neq F0(x)$
where $F0(x)$ – cumulative distribution function of a random variable $X \sim N(\mu, \sigma)$

Test statistic:

$$\boxed{\chi^2 = \sum_{i=1}^{n} \frac{(n_i - \hat{n}_i)^2}{\hat{n}_i}}$$

where:

n_i observed absolute frequencies in the interval *i*
$\hat{n}_i$ absolute frequencies in the interval *i* to be observed if $X \sim N(\mu, \sigma)$

$$\hat{n}_i = n \cdot p_i$$

Rejection region: $\chi^2 > \chi^2_{\alpha, r-k-1}$

where r-number of intervals X is grouped into, k-number of parameters estimated based on the sample

Analysis of variance (ANOVA)

Major concepts

- **Sum of squares between (SSB)** measures the variation of the sample means

$$SSB = \sum_{i=1}^{r} (\overline{x}_i - \overline{x})^2 \cdot n_i$$

- **Sum of squares for error (SSE)** measures the variation within each sample

$$SSE = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (x_{ij} - \overline{x}_i)^2$$

$$= \sum_{i=1}^{r} S_i^2 (n_i - 1)$$

- **SSB+SSE=SST** (total variation)

$$SST = S^2 \cdot (n-1)$$

Test ANOVA

H$_0$: $\mu_1 = \mu_2 = \mu_3 = \cdot\ \cdot\ \cdot = \mu_r$

H$_a$: Not all population means are equal

Test statistic:

$$F = \frac{SSB}{r-1} \cdot \frac{n-r}{SSE}$$

P-value:

$$\alpha = P(F > F_{\alpha, df1, df2})$$

where: *df1=r-1, df2=n-r*

r – number of populations compared

ANOVA table

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F	P-value
Between groups	SSB	r - 1	$\frac{SSB}{r-1}$	$F=\frac{SSB}{r-1}\cdot\frac{n-r}{SSE}$	$P(F>F_{\alpha,r-1,n-r})$
Within groups	SSE	n - r	$\frac{SSE}{n-r}$		
Total	SST	n - 1			

Two-dimensional distributions. Independence of variables. Correlation

Parameters in marginal distributions

Expected value and variance of X

$$EX = \frac{1}{n}\sum_i x_i n_{i\cdot} = \sum_i x_i p_{i\cdot}$$

$$D^2 X = \frac{1}{n}\sum_i (x_i - EX)^2 n_{i\cdot} = \sum_i (x_i - EX)^2 p_{i\cdot}$$

Expected value and variance of Y

$$EY = \frac{1}{n}\sum_j y_j n_{\cdot j} = \sum_j y_j p_{\cdot j}$$

$$D^2 Y = \frac{1}{n}\sum_j (y_j - EY)^2 n_{\cdot j} = \sum_j (y_j - EY)^2 p_{\cdot j}$$

Conditional probability

- **Conditional probability that event X=x_i occurs given that Y=y_j occurs:**

$$P(X = x_i \mid Y = y_j) = \frac{P(X = x_i \cap Y = y_j)}{P(Y = y_j)}$$

Statistical test for independence

H0: X and Y are independent
Ha: X and Y are dependent

Test statistic:

$$\boxed{\chi^2 = \sum_{ij} \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}}$$

where:
n_{ij} observed absolute frequencies
$\hat{n}_{ij}$ absolute frequencies that would be observed if X and Y independent

$$\hat{n}_{ij} = \frac{n_{i\cdot} \cdot n_{\cdot j}}{n}$$

$\hat{n}_{ij}$ must be larger than 5

Rejection region: $\chi^2 > \chi^2_{\alpha,(k-1)(l-1)}$

where *k* – number of rows and *l* – numer of columns in contingency table

Scale of the dependence

- **V-Cramer coefficient**

$$V = \sqrt{\frac{\chi^2}{n(m-1)}}$$

where m=min(k,l)

Pearson coefficient of correlation

$$r = \frac{c_{xy}}{S_x S_y}$$

- Where c_{xy} is the covariance of X and Y.
- S_x, S_y are standard deviations of X and Y respectively.

$$c_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_j - \overline{y})$$

Testing the significance of the coefficient of correlation

Two-tailed test:
H_0: ρ=0
H_1: ρ≠0

Upper-tail test:
H_0: ρ=0
H_1: ρ>0

Lower-tail test:
H_0: ρ=0
H_1: ρ<0

Test statistic:

$$t = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2}$$

P-value = P(T<-t)+P(T>t)

P-value = P(T>t)

P-value = P(T<-t)

Notations:
ρ - population coeff of correlation
r – sample coeff of corr

Spearman coefficient of correlation

$$r_d = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$

where d_i is a difference in ranks attributed to X and Y respectively

Simple regression model

Simple Regression Model

$$y = \beta_0 + \beta_1 x + \varepsilon$$

where:

y = Dependent *or* response variable (variable to be modeled)

x = Independent *or* predictor variable (variable used as a predictor of *y*)

$\beta_0 + \beta_1 x$ = Deterministic component.

ε (epsilon) = Random error component

Least Squares Estimates

SLOPE: $$\hat{\beta}_1 = \frac{\text{cov}_{xy}}{S_x^2} = r_{xy}\frac{S_y}{S_x}$$

Y-INTERCEPT: $$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

$$\text{where} \quad \text{cov}_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$$S_x^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

Estimation of σ^2

$$s_e^2 = \frac{\text{SSE}}{\text{Degrees of freedom for error}} = \frac{\text{SSE}}{n-2}$$

$$\text{where SSE} = \sum \left(y_i - \hat{y}_i \right)^2$$

$$s_e = \sqrt{s_e^2} = \sqrt{\frac{\text{SSE}}{n-2}}$$

Standard errors of βs

$$s_{\hat{\beta}_1} = \sqrt{\frac{s_e^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

$$s_{\hat{\beta}_0} = \sqrt{\frac{S_e^2 \cdot \sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

Making inferences about β

Confidence intervals:

$$P\{\hat{\beta}_i - t_{\alpha/2,n-2} S_{\hat{\beta}_i} < \beta_i < \hat{\beta}_i + t_{\alpha/2,n-2} S_{\hat{\beta}_i}\} = 1 - \alpha$$

Making inferences about β

Significance test:

Two-tailed:

$$H0: \beta_i = 0$$
$$Ha: \beta_i \neq 0$$

Lower-tail test:

$$H0: \beta_i = 0$$
$$Ha: \beta_i < 0$$

Upper-tail test:

$$H0: \beta_i = 0$$
$$Ha: \beta_i > 0$$

Test statistic: $$t = \frac{\hat{\beta}_i}{S_{\hat{\beta}_i}}$$ df=n-2

P-value = P(T<-t)+P(T>t)

P-value = P(T<-t)

P-value = P(T>t)

Verifying the Overall Utility of a Model

Coefficient of determination :

$$R^2 = \frac{\sum (\hat{y}_i - \overline{y})^2}{\sum (y_i - \overline{y})^2} = 1 - \frac{\sum (\hat{y}_i - y_i)^2}{\sum (y_i - \overline{y})^2}$$

Prediction

- Types of predictions
  - Point estimates
  - Interval estimates

- We can predict:
  - Population mean response $E(y)$ for given x *(i.e. a point* on population regression line)
  - Individual response (y_i) for given x

A 100(1 – α)% CI for the Mean Value of *y* at *x* = x_p

$$\hat{y} \pm t_{\alpha/2} s_e \sqrt{\frac{1}{n} + \frac{\left(x_p - \bar{x}\right)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$ df = $n - 2$

A 100(1 – α)% CI for New Value of *y* at *x* = x_p

$$\hat{y} \pm t_{\alpha/2} s_e \sqrt{1 + \frac{1}{n} + \frac{\left(x_p - \bar{x}\right)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$ df = $n - 2$

Basics of multiple regression

Multiple Regression Model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$$

where:

y = Dependent *or* response variable (variable to be modeled)

x = Independent *or* predictor variable (variable used as a predictor of y)

$\beta_0 + \beta_1 x$ = Deterministic component. We explain y with x

ε (epsilon) = Random error component

$y, x_1, x_2, \ldots, x_k$ are known. We use them to estimate the unknowns, the βs.

Making inferences about β

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k$$

$$[S_{\hat{\beta}_0}][S_{\hat{\beta}_1}] \quad [S_{\hat{\beta}_2}] \qquad\qquad [S_{\hat{\beta}_k}]$$

Making inferences about β

Confidence intervals:

$$P\{\hat{\beta}_i - t_{\alpha/2,n-k-1} S_{\hat{\beta}_i} < \beta_i < \hat{\beta}_i + t_{\alpha/2,n-k-1} S_{\hat{\beta}_i}\} = 1 - \alpha$$

Making inferences about β

Significance test:

Two-tailed:

$$H0:\beta_i=0$$
$$Ha:\beta_i\neq 0$$

Lower-tail test:

$$H0:\beta_i=0$$
$$Ha:\beta_i<0$$

Upper-tail test:

$$H0:\beta_i=0$$
$$Ha:\beta_i>0$$

Test statistic: $t=\dfrac{\hat{\beta}_i}{S_{\hat{\beta}_i}}$ df=n-k-1

P-value = P(T<-t)+P(T>t)

P-value = P(T<-t)

P-value = P(T>t)

Verifying the Overall Utility of a Model

Coefficient of determination :

$$R^2 = \frac{\sum(\hat{y}_i - \overline{y})^2}{\sum(y_i - \overline{y})^2} = 1 - \frac{\sum(\hat{y}_i - y_i)^2}{\sum(y_i - \overline{y})^2}$$

Time-series analysis:

moving averages, simple trend model and seasonal trend model

Odd-observation moving averages

$$_{3}\overline{y}_k = \frac{y_{k-1} + y_k + y_{k+1}}{3}$$

$$_{5}\overline{y}_k = \frac{y_{k-2} + y_{k-1} + y_k + y_{k+1} + y_{k+2}}{5}$$

Even-observation moving averages

$$ {}_2\bar{y}_k = \frac{\frac{1}{2}y_{k-1} + y_k + \frac{1}{2}y_{k+1}}{2} $$

$$ {}_4\bar{y}_k = \frac{\frac{1}{2}y_{k-2} + y_{k-1} + y_k + y_{k+1} + \frac{1}{2}y_{k+2}}{4} $$

Trend model

$$y_t = \beta_0 + \beta_1 t + \varepsilon$$

The line is fitted with LS method.

All formulas and test we learnt last time (regression lecture) hold here, i.e.
we can compute the standard errors of coeficients,
•test for significance of coefficients,
•construct confidence intervals for coefficients,
•compute the R2
•make predictions.

Seasonal regression model

$$y_t = \beta_0 + \beta_1 t + \sum_{i=1}^{k-1} \gamma_i Q_i + \varepsilon$$

Where Q_i are dummies that assume value 1 in a given subperiod and 0 otherwise

If there are k subperiods *k-1* dummies are introduced the model and teh remaining subperiod (for which the dummy is not introduced to the model) is treated as a reference category. Υ_i measure the magnitude of the seasonal fluctuation in subperiod k relative to the reference subperiod.

Index numbers

Simple index numbers

$$I_{t/t0} = \frac{Y_t}{Y_{t0}}$$

where $I_{t/t0}$ is the index number measuring a relative change in the value of Y between t_0 and t.

Average rate of change

$$i_g = \sqrt[n-1]{I_{t/t_0}} = \sqrt[n-1]{\prod_{t=1}^{n-1} I_{t/t-1}}$$

Composite index numbers

- **A weighted composite index number**

$$I_t = \frac{\sum_{i=1}^{k} Q_{it_1} P_{it_1}}{\sum_{i=1}^{k} Q_{it_0} P_{it_0}} \times 100$$

Price and quantity index numbers

Laspeyres price index:

$$I_P^L = \frac{\sum_{i=1}^{k} Q_{it_0} P_{it_1}}{\sum_{i=1}^{k} Q_{it_0} P_{it_0}}$$

Paasche price index:

$$I_P^L = \frac{\sum_{i=1}^{k} Q_{it_1} P_{it_1}}{\sum_{i=1}^{k} Q_{it_1} P_{it_0}}$$

Laspeyres quantity index:

$$I_Q^L = \frac{\sum_{i=1}^{k} Q_{it_1} P_{it_0}}{\sum_{i=1}^{k} Q_{it_0} P_{it_0}}$$

Paasche quantity index:

$$I_Q^L = \frac{\sum_{i=1}^{k} Q_{it_1} P_{it_1}}{\sum_{i=1}^{k} Q_{it_0} P_{it_1}}$$