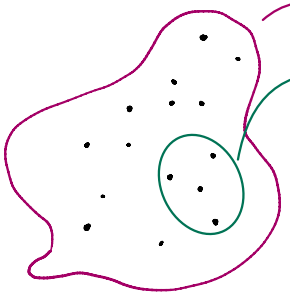
Mathematical Statistics

Notion of observations and populations



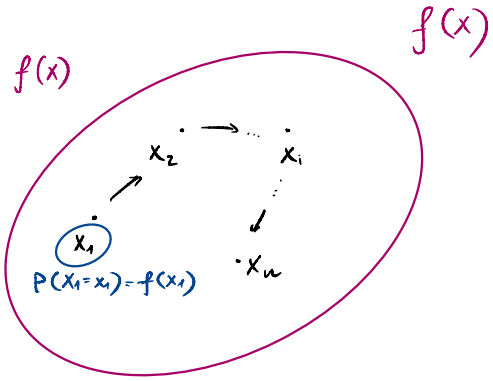
population (totality of samples) ... all students, cards in a deck

sample (observation) ... several students, one card

Each observation is a value of a random variable X having some probability distribution $f(x)$

A sample is a subset of population
Since manually chosen samples might be biased, it is worth using random sampling

In selecting random sample from population $f(x)$ let us define random variables $X_1, X_2 \dots X_i \dots X_n$ that constitute this sample with values $x_1, x_2, \dots x_i, \dots x_n$.



Because of identical conditions, variables $X_1, X_2 \dots, X_n$ are independent and indentically distributed.

So, $f(x_1, x_2 \dots, x_n) = f(x_1)\, f(x_2) \dots f(x_n)$

Our main purpose in selecting random samples is to elict information about unknown population parameters.

Any function of the random variables constituting a random sample is called statistic.

Some important statistics:

a) Sample Mean: $\frac{1}{n}\sum_{i=1}^{n} x_i = \bar{X}$

b) Sample Median: $\tilde{x} = \begin{cases} x_{(n+1)/2}, & n \text{ is odd} \\ \frac{1}{2}(x_n + x_{n+1}), & n \text{ is even} \end{cases}$

c) Sample Mode: value that occurs more often

d) Sample Variance: $\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2 = S^2$

If S^2 is sample variance of a sample of size n, then we may write:

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

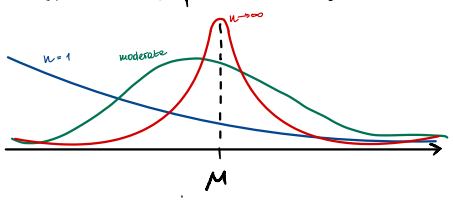
Inference about populations

- The probability distribution of a statistic is called a **sampling distribution**. The sampling distribution depends on the distribution of the population, the size of samples, and the method of choosing samples.

Sampling Distributions of Means and CLT

Central Limit Theorem. If $\bar{X}$ is a mean of a sample taken from a population with mean μ and finite variance σ^2, then the limiting form of the distribution of:

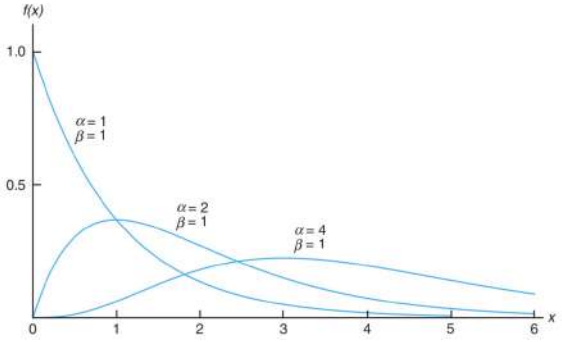
$$Z = \frac{1}{\sqrt{n}} \frac{\bar{X} - \mu}{\sigma}$$



as $n \to \infty$ is the standard normal distribution $n(z; 0; 1)$

Exercises

Gamma Distribution



- Gamma Function: $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx$
  - Properties: $\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$
    $\Gamma(1) = 1$
    $\Gamma(0.5) = \sqrt{\pi}$
- Gamma Distribution: $f_\Gamma(x;\alpha,\beta) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{\frac{-x}{\beta}} \cdot I_{x>0}$
- Exponential Distribution: $f(x;\lambda) = \lambda^{-1} e^{-\lambda x} \cdot I_{x>0}$

Chi-squared Distribution

- A variable X is said to have chi-squared distribution with ν degrees of freedom if its p.d.f is defined as: $f(x;\nu) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{\frac{\nu}{2}-1} e^{-\frac{x}{2}}$
- The mean of such distribution is ν
- The variance of such distribution is 2ν

Sampling Distribution of S^2

If S^2 is the variance of a random sample of size n taken from a normal population having the variance σ^2, then the statistic

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2}$$

has chi-squared distribution with $\nu = n-1$ degrees of freedom

t - Distribution

Let Z be a standard normal random variable and V a chi-squared random variable with ν degrees of freedom. If Z and V are independent, then the distribution of the random variable T, where

$$T = \frac{Z}{\sqrt{\nu^{-1} V}}$$

is given by the density function $h(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu}} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{1}{2}(\nu+1)}$

Let $X_1, X_2, \ldots, X_n$ be independent normally distributed random variables with mean μ and deviation σ. Let $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})$ then the random variable

$$T = \sqrt{n}\,\frac{\bar{X}-\mu}{S}$$

has a t-distribution with $\nu = n-1$ degrees of freedom.

The t-distribution is used extensively in problems with inferences about population mean or in problems that involve comparative samples

F – Distribution

Let U and V be two independent random variables having chi-squared distributions with ν_1 and ν_2 degrees of freedom. Then the distribution of the random variable

$$F = \frac{\nu_2}{\nu_1}\frac{U}{V}$$

is given by the density function

$$h(f) = \begin{cases} \dfrac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)\left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}}}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)} \dfrac{f^{\frac{\nu_1}{2}-1}}{\left[1+\frac{\nu_1}{\nu_2}f\right]^{\frac{\nu_1+\nu_2}{2}}}, & f > 0 \\ 0, & f \leq 0 \end{cases}$$

This is known as the F-distribution with ν_1 and ν_2 degrees of freedom

Writting $f_\alpha(\nu_1, \nu_2)$ for f_α with ν_1 and ν_2 degrees of freedom, we obtain:

$$f_{1-\alpha}(\nu_1, \nu_2) = \frac{1}{f_\alpha(\nu_1, \nu_2)}$$

Quantile

A quantile of a sample, $q(f)$ is a value for which a specified fraction f of the data values is less than or equal to $q(f)$

Example. If $q_X(0.8) = 4$ then $P(X \leq 4) = 0.8$

Estimation of Population Parameters

A point estimate of some population parameter θ is a single value $\hat{\theta}$ of a statistic $\hat{\Theta}$.

A statistic $\hat{\Theta}$ is said to be an unbiased estimator for a parameter θ if

$$\mu_{\hat{\theta}} = E(\hat{\Theta}) = \theta$$

Estimator biasness

Gorodetskiy's Notes

$$\text{bias}\,\hat{\theta} = E\hat{\Theta} - \theta$$

Mean Square Error

$$E(\hat{\theta}(\vec{x}) - \theta)^2 = \text{Var}\,\hat{\theta}(\vec{x}) + \text{bias}^2\hat{\theta}(\vec{x})$$

Consistency

Estimator $\hat{\theta}$ is said to be consistent if $\forall \varepsilon > 0 \quad \lim_{n\to\infty} P(|\hat{\theta} - \theta| \geq \varepsilon) = 0$

$$P(|\hat{\theta} - \theta| \geqslant \varepsilon) \leq \frac{\text{var}\,\hat{\theta}}{\varepsilon^2} \longrightarrow 0$$

($E\hat{\theta}$ (unbiased))

Variance of Estimators

If we consider all unbiased estimators of some population parameter θ, the one with the smallest variance is called the most efficient estimator of θ.

Interval Estimation

There are many situations in which it is preferable to determine an interval within we would expect to find the value of the parameter Such an interval is called interval estimate. Such estimate can be expressed as two numbers θ_L and θ_u such that $\theta_L \leq \theta \leq \theta_u$

Estimating the mean

Recalling Central Limit Theorem, we can obtain a confidence interval for a sample mean estimation.

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1-\alpha$$

where

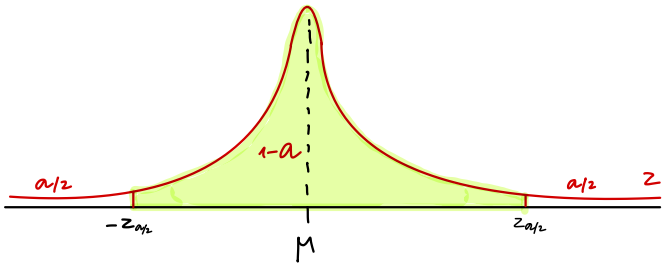
$$Z = \sqrt{n}\,\frac{\bar{X}-\mu}{\sigma}$$

$$\Phi_0(x) = \int_0^x \frac{e^{-t^2/2}}{\sqrt{2\pi}}\,dt = \frac{\alpha}{2}$$

$$x = \Phi_0^{-1}\left(\frac{\alpha}{2}\right)$$

We obtain: ($\Phi_0^{-1}(\frac{\alpha}{2})$)

$$P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$



If $\bar{x}$ is the mean of a random sample of size n from a population with known variance σ^2, a $100(1-\alpha)\%$ confidence interval for μ is given by:

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Example Exercise I.

The average zinc concentration recovered from a sample of measurements taken in 36 different locations in a river is found to be 2.6 g. per mm. Find the 95% and 99% confidence intervals for the mean zinc concentration. Assume that population std. deviation is 0.3 g. per mm

a) ($z_{0.025}$, $\bar{x}$, σ, $\sqrt{n}$)

$$2.6 - (1.96)\left(\frac{0.3}{\sqrt{36}}\right) < \mu < 2.6 + (1.96)\left(\frac{0.3}{\sqrt{36}}\right)$$

$$2.5 < \mu < 2.7$$

b)

$$2.6 - (2.575)\left(\frac{0.3}{\sqrt{36}}\right) < \mu < 2.6 + (2.575)\left(\frac{0.3}{\sqrt{36}}\right)$$

$$2.47 < \mu < 2.73$$

If $\bar{x}$ is used as an estimate of μ, we can be $100(1-\alpha)\%$ confident that the error will not exceed a specified amount e when the sample size is

$$n = \left(\frac{z_{\alpha/2}\,\sigma}{e} \right)^2$$

Example Exercise 11:

How large a sample is required if we want to be 95% confident that our estimate of μ in previous exercise is off by less than 0.05.

$$n = \left[\frac{1.96 \cdot 0.3}{0.05} \right]^2 = 138.3 \longrightarrow 139$$

One-sided intervals

$$P\left(\mu > \bar{x} - z_\alpha \frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

$$P\left(\mu < \bar{x} + z_\alpha \frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

Large-Sample Confidence Interval

Often even when normality cannot be assumed, if $n \geq 30$, s can replace σ and the confidence interval

$$\bar{x} = \pm z_{\alpha/2} \frac{s}{\sqrt{n}}$$

may be used

The case of σ unknown

If the particular value of σ is unknown, it can be replaced with statistic S. Then, recalling that statistic (considering normal sample)

$$T = \sqrt{n}\,\frac{\bar{X}-\mu}{S}$$

has a t-distribution with $n-1$ degrees of freedom. Hence, we can use that fact in the same manner as we used CLT.

$$P(-t_{\alpha/2} < T < t_{\alpha/2}) = 1-\alpha$$

$$P\left(-t_{\alpha/2} < \sqrt{n}\,\frac{\bar{X}-\mu}{S} < t_{\alpha/2}\right) = 1-\alpha$$

$$P\left(\bar{X} - t_{\alpha/2}\frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{\alpha/2}\frac{S}{\sqrt{n}}\right) = 1-\alpha$$

Standart Error of Point Estimate

A standart error of a point estimate of $\bar{X}$ is given as $\frac{\sigma}{\sqrt{n}}$

Prediction Intervals

For a normal distribution of measurements with unknown mean μ and known variance σ^2, a $100(1-\alpha)\%$ prediction interval of a future observation x_0 is:

$$\bar{x} - z_{\alpha/2}\,\sigma\sqrt{1+\tfrac{1}{n}} < x_0 < \bar{x} + z_{\alpha/2}\,\sigma\sqrt{1+\tfrac{1}{n}}$$

where $z_{\alpha/2}$ is the z-value leaving an area of $\alpha/2$ to the right.
or

$$\bar{x} - t_{\alpha/2}\,s\sqrt{1+\tfrac{1}{n}} < x_0 < \bar{x} + t_{\alpha/2}\,s\sqrt{1+\tfrac{1}{n}}$$

where $t_{\alpha/2}$ is the t-value leaving an area of $\alpha/2$ to the right

Tolerance Intervals

For a normal distribution of measurements with unknown mean μ and unknown std. deviation σ, tolerance limits are given by $\bar{x} \pm ks$, where k is determined such that one can assert with $100(1-\gamma)\%$ confidence that the given limits contain at least the proportion $1-\alpha$ of the measurements. (Table values)

Estimating the differences between two means

If we have two populations with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2 respectively, a point estimator of the difference between μ_1 and μ_2 is given by the statistic $\bar{X}_1 - \bar{X}_2$ that can be approximated using the normal distribution

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Estimating the Variance

An interval estimate of σ^2 can be established by using the statistic:

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

If samples are taken from normal opulation, then $X^2 \sim \chi^2_{n-1}$

So,

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

Sufficient Statistic

The statistic $T(\vec{x})$ is said to be sufficient for parameter θ if $\forall D \in \mathbb{R}^n \; P(\vec{x} \in D \mid T(\vec{x}))$ is not dependent on θ.

$T(\vec{x})$ is sufficient for θ if

$$f_{\vec{X}}(\vec{x}, \theta) = g(\vec{x})\, h(T(\vec{x}), \theta) \quad \text{(factorization criterion)}$$

Example

1) $\bar{X}$ is sufficient for θ if $X_1, X_2 \ldots, X_n \sim \text{Pois}(\theta)$

$$P(X=k) = e^{-\theta}\frac{\theta^k}{k!}$$

$$f_{\vec{X}}(\vec{x}, \theta) = \prod_{i=1}^{n} e^{-\theta}\frac{\theta^{x_i}}{x_i!} = \underbrace{\frac{1}{\prod_{i=1}^{n} x_i!}}_{g(\vec{x})} \underbrace{e^{-n\theta}\cdot\theta^{n\bar{x}}}_{h(\vec{x},\theta)}$$

Method of maximum likelihood

Likelihood function: $L(x_1, \dots, x_n, \theta) = f(x_1, \dots, x_n, \theta) = \prod_{i=1}^{n} f_i(x_i, \theta)$

The likelihood of the sample is the following joint probability:

$$P(X_1 = x_1, \dots, X_n = x_n \mid \theta)$$

or

$$P(X_1 = x_1 \mid \theta) \dots P(X_n = x_n \mid \theta)$$

This function is ought to be maximized

Examples

I) Suppose $X_1, X_2 \dots X_n$ are i.i.d random variables distributed as $N(0,1)$. Find good estimators for parameters μ and σ^2

$$L(\vec{x}; \mu; \sigma) = \prod_{x_i \in \vec{x}} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(x_i-\mu)^2\right)$$

Likelihood function

Since $\ln$ is a monotone increasing function it is possible to maximize $\ln L$ to acquire maximum for L

$$\ln L = -\frac{n}{2}\ln(2\pi) - n\ln\sigma - \frac{1}{2\sigma^2}\sum_{j=1}^{n}(x_j-\mu)^2$$

$$\begin{cases} L_\mu = \sigma^{-2}\sum_{j=1}^{n}(x_j-\mu) = 0 \\ L_\sigma = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{j=1}^{n}(x_j-\mu)^2 = 0 \end{cases}$$

$$\begin{cases} \frac{1}{n}\sum_{j=1}^{n} x_j = \mu = \bar{x} \\ \sigma^2 = \frac{1}{n}\sum_{j=1}^{n}(x_j-\bar{x})^2 \end{cases} \quad \Bigg| \quad \left.\begin{matrix} \hat{\mu} = \bar{x} \\ \sigma^2 = s^2 \end{matrix}\right\} \text{good estimators for } N(0;1)$$

Tests of Hypotheses

A statistical hypothesis is an assertion or conjecture concerning one or more populations.

The structure of hypothesis testing will be formulated with the use of the term null hypothesis, which refers to any hypothesis we wish to test and is denoted as H_0. The rejection of H_0 leads to acceptance of H_1 (alternative hypothesis)

Errors

Type I Error: Rejection of the null hypothesis when it is true.

Type II Error: Nonrejection of the null hypothesis when it is false

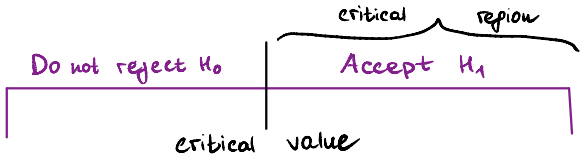
The probability of committing a type I error is called the level of significance, and is denoted via Greek letter α

The probability of committing a type II error is denoted via letter β

- The type I error and type II error are related. A decrease in the probability of one generally results
- The size of the critical region and therefore the probability of committing a type I error, can always be reduced by adjusting the critical values.
- An increase in the sample size n will reduce both α and β
- The greater the distance between the true value and hypothesized value, the smaller β will be.

The power of a test is the probability of rejecting H_0 given that alternative is true.

The power of a test may be computed as $1-\beta$



Tests Concerning Sample Mean

Suppose $X_1, X_2, \dots, X_n$ represent a random sample from a distribution with mean μ and variance $\sigma^2 > 0$.

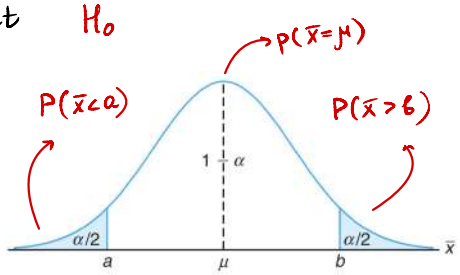
$H_0: \mu = \mu_0$
$H_1: \mu \neq \mu_0$

Using the property of standard normal distribution we can derive a critical region for rejecting hypothesis H_0 with probability α to reject it when its true

If $-z_{\alpha/2} < z = \sqrt{n}\frac{\bar{x} - \mu_0}{\sigma} < z_{\alpha/2}$ do not reject H_0

or reject if

$\bar{x} < \mu_0 - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ or $\bar{x} > \mu_0 + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$



Test on Two Means

The two-sided hypothesis on two means can be written generally as

$$H_0: \mu_1 - \mu_2 = d_0$$

The test-statistics is given as:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Reject H_0 if $z > z_{\alpha/2}$ or $z < -z_{\alpha/2}$

Pooled t-test

$H_0: \mu_1 = \mu_2$
$H_1: \mu_1 \neq \mu_2$ (assume $\sigma_1 = \sigma_2$)

Test statistics: $t = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{s_p\sqrt{n_1^{-1} + n_2^{-1}}}$

$s_p = \frac{s_1^2(n-1) + s_2^2(n-1)}{n_1 + n_2 - 2}$

Reject H_0 if $s_p^2 > t_{\alpha/2, n_1+n_2-2}$ or $s_p^2 < -t_{\alpha/2, n_1+n_2-2}$

Unknown and Unequal Variances

The test statistic is given as

$$T' = \frac{(\bar{X}_1 - \bar{X}_2) - d_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

and has $\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{s_1^2}{n_1}\right)^2(n_1-1)^{-1} + \left(\frac{s_2^2}{n_2}\right)^2(n_2-1)^{-1}}$ degrees of freedom

As a result, we do not reject H_0 when $-t_{\alpha/2,\nu} < t' < t_{\alpha/2,\nu}$

Table 10.3: Tests Concerning Means

H_0	Value of Test Statistic	H_1	Critical Region
$\mu = \mu_0$	$z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$; σ known	$\mu < \mu_0$ $\mu > \mu_0$ $\mu \neq \mu_0$	$z < -z_\alpha$ $z > z_\alpha$ $z < -z_{\alpha/2}$ or $z > z_{\alpha/2}$
$\mu = \mu_0$	$t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$; $v = n - 1$, σ unknown	$\mu < \mu_0$ $\mu > \mu_0$ $\mu \neq \mu_0$	$t < -t_\alpha$ $t > t_\alpha$ $t < -t_{\alpha/2}$ or $t > t_{\alpha/2}$
$\mu_1 - \mu_2 = d_0$	$z = \dfrac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$; σ_1 and σ_2 known	$\mu_1 - \mu_2 < d_0$ $\mu_1 - \mu_2 > d_0$ $\mu_1 - \mu_2 \neq d_0$	$z < -z_\alpha$ $z > z_\alpha$ $z < -z_{\alpha/2}$ or $z > z_{\alpha/2}$
$\mu_1 - \mu_2 = d_0$	$t = \dfrac{(\bar{x}_1 - \bar{x}_2) - d_0}{s_p\sqrt{1/n_1 + 1/n_2}}$; $v = n_1 + n_2 - 2$, $\sigma_1 = \sigma_2$ but unknown, $s_p^2 = \dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$	$\mu_1 - \mu_2 < d_0$ $\mu_1 - \mu_2 > d_0$ $\mu_1 - \mu_2 \neq d_0$	$t < -t_\alpha$ $t > t_\alpha$ $t < -t_{\alpha/2}$ or $t > t_{\alpha/2}$
$\mu_1 - \mu_2 = d_0$	$t' = \dfrac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$; $v = \dfrac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$, $\sigma_1 \neq \sigma_2$ and unknown	$\mu_1 - \mu_2 < d_0$ $\mu_1 - \mu_2 > d_0$ $\mu_1 - \mu_2 \neq d_0$	$t' < -t_\alpha$ $t' > t_\alpha$ $t' < -t_{\alpha/2}$ or $t' > t_{\alpha/2}$
$\mu_D = d_0$ paired observations	$t = \dfrac{\bar{d} - d_0}{s_d/\sqrt{n}}$; $v = n - 1$	$\mu_D < d_0$ $\mu_D > d_0$ $\mu_D \neq d_0$	$t < -t_\alpha$ $t > t_\alpha$ $t < -t_{\alpha/2}$ or $t > t_{\alpha/2}$

Goodness-of-fit test

This kind of the test is ought to determine the distribution of the given population

We suppose that given population has a certain distribution and therefore sampling vector is expected as $\vec{o} = (o_1, o_2, \dots, o_n)^T$ where o_i is expected number of occurences of i-th outcome.

Suppose $\vec{e}$ is a vector of observed number of occurences.
Then

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

approx. has chi-squared distribution with $k-1$ degrees of freedom.

Test for Independence

χ^2 test can be used for testing hypothesis of independence
Let our observations consists of pairs of i.i.d.r.v $(X_1; Y_1), (X_2; Y_2) \dots (X_k; Y_k)$

H_0: X_i and Y_i are independent

$x \backslash y$	b_1	b_2	.....	b_n	Σ	
a_1	ϑ_{11}	ϑ_{12}		ϑ_{1n}	ϑ_{1+}	p_{1+}
a_2	ϑ_{21}			ϑ_{2n}	ϑ_{2+}	p_{2+}
$\vdots$						
a_m	ϑ_{m1}			ϑ_{mn}		p_{m+}
Σ	ϑ_{+1}	ϑ_{+2}		ϑ_{+n}	ϑ_{++}	
	p_{+1}	p_{+2}		p_{+n}		

$\forall i,j: \ 1 \le i \le m \quad 1 \le j \le n \quad P(X = a_i, Y = b_j) = P(X = a_i) P(Y = b_j)$
independence condition

Let ϑ_{ij} be a number of outcomes such that $\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} a_i \\ b_j \end{pmatrix}$

H_0: $P(X = a_i, Y = b_j) = p_{i+} \cdot p_{+j} \quad \forall i,j$

1) Using the method of maximum likelihood: $L = \prod_{\substack{1 \le i \le m \\ 1 \le j \le n}} (p_{i+} p_{+j})^{\vartheta_{ij}} = \prod_{1 \le i \le m} p_{i+}^{\vartheta_{i+}} \prod_{1 \le j \le n} p_{+j}^{\vartheta_{+j}}$

$\ln L = \sum_{i=1}^{m} \vartheta_{i+} \ln p_{i+} + \sum_{j=1}^{n} \vartheta_{+j} \ln p_{+j}$

under condition $p_{1+} + p_{2+} + \dots + p_{m+} = 1$

$p_{i+} = \frac{\vartheta_{i+}}{\vartheta_{++}}$

$p_{j+} = \frac{\vartheta_{+j}}{\vartheta_{++}}$

$\tilde{p}_{ij} = \frac{\vartheta_{i+} \vartheta_{+j}}{\vartheta_{++}^2}$

$$\sum \frac{\left(\vartheta_{ij} - \frac{\vartheta_{i+}\vartheta_{+j}}{\vartheta_{++}^2}\right)^2}{\frac{\vartheta_{i+}\vartheta_{+j}}{\vartheta_{++}^2}\vartheta_{++}} = \vartheta_{++} \sum_{ij} \frac{\left(\vartheta_{ij} - \frac{\vartheta_{i+}\vartheta_{+j}}{\vartheta_{++}}\right)}{\vartheta_{i+}\vartheta_{+j}} \sim \chi^2_{(n-1)(m-1)}$$

2) Calculate the sum and decide about the hypothesis.

Linear Regression

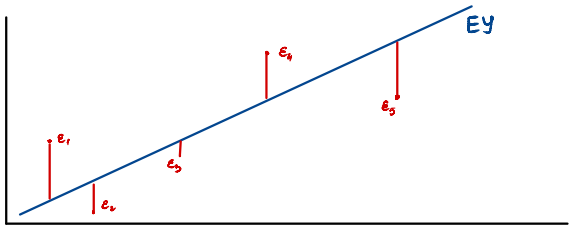
Simple Linear Regression Model

The model must include the set $\{(x_i, y_i): i = 1, 2, \dots, n\}$ of data involving n pairs of (xy) values. The value y_i depends on x_i via a linear structure that also has the random component involved. The basis for the use of a statistical model relates to how the random variable Y moves with X and the random variable.

$$Y = \beta_0 + \beta_1 x + \varepsilon$$ → random variable (error)

where β_0 and β_1 are unknown intercept and slope, while ε is a random variable that is distributed with $E(\varepsilon) = 0$ and $\text{var}(\varepsilon) = \sigma^2$

Since $E(\varepsilon)$ is assumed to be 0, then $EY = \beta_0 + \beta_1 x$ is a regression line



The purpose of the regression analysis is to estimate unknown coefficients β_0 and β_1

The estimated regression line is given as: $\hat{y} = b_0 + b_1 x$

Residual: Error in Fit

Give a set of a regression data $\{(x_i, y_i); i = 1, 2, \dots, n\}$ and a fitted model $\hat{y}_i = b_0 + b_1 x_i$, the i-th residual is given as $e_i = y_i - \hat{y}_i$, $i = 1, 2 \dots n$

The Method of Least Squares

$$SSE = \sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x)^2$$

Differentiating SSE yields:

$$SSE_{b_0} = -2\sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) \qquad SSE_{b_1} = -2\sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) x_i$$

Setting both derivatives to 0, we obtain values for b_0 and b_1:

$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \quad \text{and} \quad b_0 = \bar{y} - b_1 \bar{x}$$

Properties of Estimators

Denote $R = \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta}(x_i - \bar{x}))^2$ as a residual sum of squares and $\varepsilon_i \sim N(0, \sigma^2)$

The following statements hold:

1) $\hat{\alpha} \sim N(\alpha, \frac{\sigma^2}{n})$, $\hat{\beta} \sim N(\beta, \frac{\sigma^2}{S_{xx}})$
2) $\frac{R}{\sigma^2} \sim \chi^2_{n-2}$
3) $\hat{\alpha}$, $\hat{\beta}$, R are independent
4) $(n-2)^{-1} R$ is an unbiased estimator for σ^2