Formelsamling för tenta i Dataanalys och Statistik, tms136, I2.

Combinatorics:

 $\binom{n}{r} = \frac{n!}{r!(n-r)!} = \text{number of subsets of size r from a set of size n}$ $n! = n(n-1)(n-2) \cdot ... \cdot 2 \cdot 1 = \text{number of permutations of n different}$ objects.

Events

 $P(B) = P(B \cap A) + P(B \cap A') = P(B|A)P(A) + P(B|A')P(A')$ $P(A \cap B) = P(B|A)P(A) = P(A|B)P(B)$ $P(A \cup B) = P(A) + P(B) - P(A \cap B) \le 1$ $E_1 \cap E_2 = E_1 \cap E_1' = 0$ – Mutually exclusive

Probability mass/density function:

- $(1) \quad f(x) \ge 0$
- (2) $\sum_{i=1}^{n} f(x_i) = \int_{-\infty}^{+\infty} f(x) dx = 1$
- (3) $f(x_i) = P(X = x_i)$, for discrete variables
- (4) $P(a \le X \le b) = \int_a^b f(x) dx$, for continuous variables

Cumulative distribution function:

$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(u)du, \quad for -\infty < x < \infty$$

Mean and variance of a random variable:

$$\mu = E(X) = \sum_{x} x f(x) = \int_{-\infty}^{\infty} x f(x) dx$$

$$\sigma^{2} = Var(X) = \sum_{x} x^{2} f(x) - \mu^{2} = \int_{-\infty}^{\infty} x^{2} f(x) dx - \mu^{2} =$$

$$= E(X^{2}) - (E(X))^{2}$$

Expected value of a function of a continuous random value:

$$E[h(x)] = \int_{-\infty}^{\infty} h(x)f(x)dx$$

$$f(x_i) = \frac{1}{x}$$
, for all x

Discrete uniform distribution,
$$Uni(\{x_1, ..., x_n\})$$

$$f(x_i) = \frac{1}{n}, for \ all \ x_i$$
Continuous uniform distribution, $Uni([a, b])$

$$f(x) = \frac{1}{(b-a)}, E(X) = \frac{(a+b)}{2}, Var(X) = \frac{(b-a)^2}{12}$$
Binomial distribution, $Bin(n, p)$

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

$$E(X) = np, \qquad Var(x) = np(1-p)$$

Geometric distribution, Geo(p)

$$f(x) = (1-p)^{x-1}p, \qquad x = 1,2,...$$

$$\mu = E(X) = \frac{1}{p}, \qquad \sigma^2 = Var(x) = (1-p)/p^2$$

Poisson distribution, $Poi(\lambda)$

$$f(x) = \frac{e^{-\lambda}\lambda^x}{x!}, x = 0, 1, 2, ...$$

$$\mu = E(X) = Var(X) = \lambda$$

Normal distribution,
$$N(\mu, \sigma^2)$$

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

$$E(X) = \mu, \quad Var(x) = \sigma^2$$
Exponential distribution. Exponential

$$E(X) = \mu$$
, $Var(x) = \sigma^2$

Exponential distribution, $Exp(\lambda)$:

$$f(x) = \lambda e^{-\lambda x}$$
, for $0 \le x \le \infty$
 $\mu = E(X) = 1/\lambda$, $\sigma^2 = Var(X) = 1/\lambda^2$

$$Z = \frac{X - np}{\sqrt{np(1-p)}}$$

is approximately a standard normal random variable.

The approximation is good for np > 5 and n(1-p) > 5

Normal approximation to the Poisson distribution:

$$Z = \frac{x-\lambda}{\sqrt{\lambda}}$$
 is good for $\lambda > 5$

Joint probability mass function of two random variables:

- $(1) \quad f_{XY}(x,y) \geq 0$

- (1) $f_{XY}(x,y) \ge 0$ (2) $\sum_{x} \sum_{y} f_{XY}(x,y) = \iint_{-\infty}^{\infty} f_{XY}(x,y) dx dy = 1$ (3) $f_{XY}(x,y) = P(X = x, Y = y)$ (discrete case) (4) $P(X < x, Y < y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{XY}(x,y) dy dx$ (continuous case)

Marginal probability mass function:

$$f_X(x) = P(X = x) = \sum_y f_{XY}(x, y)$$
 (discrete case)

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y) dy$$
 (continuous case)

Conditional probability mass/density function of Y given X=x is:

$$f_{Y|x}(y) = \frac{f_{XY}(x,y)}{f_X(x)}, \quad for f_X(x) > 0$$

Covarience:

$$Cov(X,Y) = \sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)] = E(XY) - \mu_X \mu_Y$$

Correlation:

$$\rho_{XY} = \frac{Cov(X, Y)}{\sqrt{V(X)V(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}, \quad -1 \le \rho_{XY} \le 1$$

If X & Y are independent random variables, then $Cov(X,Y) = \rho_{XY} = 0 \ (\Leftarrow)$

The **standard error** of an estimator $\widehat{\Theta}$ is its standard deviation, given by:

$$\sigma_{\widehat{\Theta}} = \sqrt{V(\widehat{\Theta})}$$

The **mean squared error** of the estimator $\widehat{\Theta}$ of the parameter θ is:

$$MSE(\widehat{\Theta}) = E(\widehat{\Theta} - \theta)^2$$

An estimator $\widehat{\Theta}$ is called **unbiased** if

$$E(\hat{\theta}) - \theta = 0$$

Likelihood function:

$$L(\theta) = \prod_{i=1}^{n} f(x_i|\theta) = f(x_1|\theta)f(x_2|\theta) \cdots f(x_n|\theta)$$

If x_1, \dots, x_n is a sample of n observations, the **sample variance** is:

$$s^{2} = \frac{\sum_{i=1}^{n} (x_{i} - \bar{x})^{2}}{n - 1} = \frac{\sum_{i=1}^{n} x_{i}^{2} - \frac{(\sum_{i=1}^{n} x_{i})^{2}}{n}}{n - 1}$$
Confidence integral on the most variance less than the many variance less than t

Confidence interval on the mean, variance known:

$$\overline{c} - z_{\alpha/2} \sigma / \sqrt{n} \le \mu \le \overline{x} + z_{\alpha/2} \sigma / \sqrt{n}$$

$$\begin{split} &\bar{x} - z_{\alpha/2} \sigma / \sqrt{n} \leq \mu \leq \bar{x} + z_{\alpha/2} \sigma / \sqrt{n} \\ &Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \,, \qquad \textit{Choice of } n = \left(\frac{z_{\alpha/2} \sigma}{E}\right)^2, E = |\bar{x} - \mu| \end{split}$$

Confidence interval on the mean, variance unknown:

$$\bar{x} - t_{\alpha/2, n-1} s / \sqrt{n} \le \mu \le \bar{x} + t_{\alpha/2, n-1} s / \sqrt{n}$$

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

Random sample normal disr. mean= μ , var= σ^2 , S²=sample var.

$$\chi_{n-1}^2 = \frac{(n-1)S^2}{\sigma^2}$$
 has χ^2 dist. with $n-1$ degrees of freedom CI on variance, s^2 =sample variance, σ^2 unknown σ^2

$$\frac{(n-1)s^2}{\chi_{\frac{\alpha}{2},n-1}^2} \le \sigma^2 \le \frac{(n-1)s^2}{\chi_{1-\frac{\alpha}{2},n-1}^2}$$

Lower and upper confidence bounds on
$$\sigma^2$$
:
$$\frac{(n-1)s^2}{\chi^2_{\alpha,n-1}} \leq \sigma^2 \text{ and } \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{1-\alpha,n-1}}$$

Proportion:

If n is large, the distribution of

$$Z = \frac{X - np}{\sqrt{np(1-p)}} = \frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

is approximately standard normal. CI on proportion (obs, lower, upper change $z_{\alpha/2}$ to z_{α} :

$$\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le p \le \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$
Sample size for a specified error on binomial proportion:

$$n = \left(\frac{Z\alpha}{\frac{2}{E}}\right)^2 p(1-p), n \text{ is } \max for \ p = 0.5$$

CI, difference in mean, variances known:

$$\bar{x}_1 - \bar{x}_2 - z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq \bar{x}_1 - \bar{x}_2 + z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

for one-sided, change $z_{\alpha/2}$ to $z_\alpha.$ Sample size for a CI on difference in mean, variances known:

$$n = \left(\frac{Z_{\alpha/2}}{F}\right)^2 \left(\sigma_1^2 + \sigma_2^2\right)$$

 $n = \left(\frac{Z_{\alpha/2}}{E}\right)^2 (\sigma_1^2 + \sigma_2^2)$ CI Case 1, difference in mean, variance unknown & equal:

$$\begin{split} \bar{x}_1 - \bar{x}_2 - t_{\alpha/2, n_1 + n_2 - 2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} &\leq \mu_1 - \mu_2 \\ &\leq \bar{x}_1 - \bar{x}_2 + t_{\alpha/2, n_1 + n_2 - 2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \end{split}$$

$$S_p^2=\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$$
 CI Case 2, difference in mean, variance unknown, not equal:

$$\begin{split} & \bar{x}_1 - \bar{x}_2 - t_{\alpha/2,\nu} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq \bar{x}_1 - \bar{x}_2 + t_{\alpha/2,\nu} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \\ & \nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}} \\ & \nu \text{ is degrees of freedom for } t_{\alpha/2,\text{if not integer, round down.} \end{split}$$

CI for μ_D from paired samples:

$$\bar{d} - t_{\alpha/2, n-1} s_D / \sqrt{n} \le \mu_D \le \bar{d} + t_{\alpha/2, n-1} s_D / \sqrt{n}$$

Approximate CI on difference in population proportions:

$$\begin{split} \hat{p}_1 - \hat{p}_2 - z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}} &\leq p_1 - p_2 \\ &\leq \hat{p}_1 - \hat{p}_2 + z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}} \end{split}$$

Hypothesis test:

- Choose parameter of interest
- H₀:
- 3. H₁:
- The test statistic is
- Reject H_0 at $\alpha = ...$ if
- 7. Computations8. Conclusions

Test on mean, variance known

$$H_0$$
: $\mu = \mu_0$, $Z_0 = \frac{X - \mu_0}{\sigma / \sqrt{n}}$

Alternative hypothesis	Rejection criteria
H_1 : $\mu \neq \mu_0$	$z_0 > z_{\alpha/2}$ or $z_0 < -z_{\alpha/2}$
H_1 : $\mu > \mu_0$	$z_0 > z_{\alpha}$
H_1 : $\mu < \mu_0$	$z_0 < -z_\alpha$

Test on mean, variance unknown

$$H_0$$
: $\mu=\mu_0$, $T_0=rac{\overline{X}-\mu_0}{S/\sqrt{n}}$

Alternative hypothesis	Rejection criteria
H ₁ : μ ≠ μ ₀	$t_0 > t_{\alpha/2, n-1} \text{ or } t_0 < -t_{\alpha/2, n-1}$
H_1 : $\mu > \mu_0$	$t_0 > t_{\alpha,n-1}$
H_1 : $\mu < \mu_0$	$t_0 < -t_{\alpha, n-1}$

Test in the variance of a normal distribution:

$$H_0: \sigma^2 = \sigma_0^2, \qquad \chi_0^2 = \frac{(n-1)S^2}{\sigma_0^2}$$

	σ_0^2	
Alternative hypothesis		Rejection criteria
$H_1: \sigma^2 \neq \sigma_0^2$		$\chi_0^2 > \chi_{\alpha/2,n-1}^2 \text{ or } \chi_0^2 < \chi_{1-\alpha/2,n-1}^2$
$H_1: \sigma^2 > \sigma_0^2$		$\chi_0^2 > \chi_{\alpha,n-1}^2$
$H_1: \sigma^2 < \sigma_0^2$		$\chi_0^2 < \chi_{1-\alpha,n-1}^2$

Approximate test on a proportion:
$$H_0 \colon p = p_0, \qquad Z_0 = \frac{X - np_0}{\sqrt{np_0(1 - p_0)}}$$

	$\sqrt{np_0(1-p_0)}$		
Alternative hypothesis		Rejection criteria	(**)
H ₁ : p ≠ p ₀		$z_0 > z_{\alpha/2}$ or $z_0 < -z_{\alpha/2}$	
$H_1: p > p_0$		$z_0 > z_\alpha$	
$H_1: p < p_0$		70 < -7	

App. Sample size for a 2-sided test on a proportion:
$$n = \begin{bmatrix} z_{\alpha/2} \sqrt{p_0(1-p_0)} + z_\beta \sqrt{p(1-p)} \\ p-p_0 \end{bmatrix}^2, for \ 1-sided \ use \ z_\alpha$$

$$H_0$$
: $\mu_1 - \mu_2 = \Delta_0$, $Z_0 = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$

Alternative hypothesis	Rejection criteria
H_1 : $\mu_1 - \mu_2 \neq \Delta_0$	$z_0 > z_{\alpha/2}$ or $z_0 < -z_{\alpha/2}$
$H_1: \mu_1 - \mu_2 > \Delta_0$	$z_0 > z_\alpha$
$H_1: \mu_1 - \mu_2 < \Delta_0$	$z_0 < -z_\alpha$

Sample size, 1-sided test on difference in mean, with power of at least 1-β,

$$n = \frac{(z_{\alpha} + z_{\beta})^{2} (\sigma_{1}^{2} + \sigma_{2}^{2})}{(\Delta - \Delta_{0})^{2}}$$

$$\begin{split} &\mathbf{n}_1 = \mathbf{n}_2 = \mathbf{n}, \text{ variance known:} \\ &n = \frac{\left(z_\alpha + z_\beta\right)^2 (\sigma_1^2 + \sigma_2^2)}{(\Delta - \Delta_0)^2} \\ &\textbf{Tests on diff. in mean, } \text{ variances unknown and equal:} \\ &H_0 : \mu_1 - \mu_2 = \Delta_0, \qquad &T_0 = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \end{split}$$

Alternative hypothesis	Rejection criteria
$H_1: \mu_1 - \mu_2 \neq \Delta_0$	$t_0 > t_{\alpha/2, n_1 + n_2 - 2} \ or$
	$t_0 < -t_{\alpha/2, n_1 + n_2 - 2}$
H_1 : $\mu_1 - \mu_2 > \Delta_0$	$t_0 > t_{\alpha, n_1 + n_2 - 2}$
$H_1:\ \mu_1-\mu_2<\Delta_0$	$t_0 < -t_{\alpha, n_1 + n_2 - 2}$

Tests on diff. in mean, variances unknown and not equal:

If
$$H_0$$
: $\mu_1 - \mu_2 = \Delta_0$ is true, the statistic $T_0 = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$

is distributed app. as t with v degrees of freedom, $\sim t(v)$

Paired t-test:

$$\begin{aligned} & \textbf{Paired r-test:} \\ & H_0 \colon \mu_D = \mu_1 - \mu_2 = \Delta_0, \qquad T_0 = \frac{\overline{D} - \Delta_0}{S_D / \sqrt{n}}, \qquad d = \frac{\mu_D}{\sigma_D} = \frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2 + \sigma_2^2}} \\ & \underline{\textbf{Alternative hypothesis}} & \textbf{Rejection criteria} \\ & \underline{\textbf{H}_1 \colon \mu_D \neq \Delta_0} & \textbf{t}_0 > \textbf{t}_{\alpha/2, n-1} \text{ or } \textbf{t}_0 < -\textbf{t}_{\alpha/2, n-1} \\ & \underline{\textbf{H}_1 \colon \mu_D > \Delta_0} & \textbf{t}_0 > \textbf{t}_{\alpha, n-1} \\ & \underline{\textbf{H}_1 \colon \mu_D < \Delta_0} & \textbf{t}_0 < -\textbf{t}_{\alpha, n-1} \end{aligned}$$

Approximate tests on the difference of two population proportions:

$$H_0: p_1 = p_2, \ Z_0 = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{\hat{P}(1-\hat{P})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \hat{P} = \frac{X_1 + X_2}{n_1 + n_2}, see \ (**) \uparrow$$

Expected frequency:
$$E_i = np_i$$
, $p_i = P(X = x) = f(x)$
The power of a test: $e^{-kx} = 1 - \beta$

$$\beta = \Phi \left(z_{\alpha/2} - \frac{\Delta - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}\right) - \Phi \left(-z_{\alpha/2} - \frac{\Delta - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}\right)$$

The P-value is the smallest level of significance that would lead to rejection of the null hypothesis H₀ with the given data.

$$P = \begin{cases} 2[1 - \Phi(|z_0|)] \text{ for a two - tailed test: } H_o: \mu = \mu_0 & H_1: \mu \neq \mu_0 \\ 1 - \Phi(z_0) & \text{for a upper - tailed test: } H_o: \mu = \mu_0 & H_1: \mu > \mu_0 \\ \Phi(z_0) & \text{for a lower - tailed test: } H_o: \mu = \mu_0 & H_1: \mu < \mu_0 \end{cases}$$

the null hypothesis H₀ with the given data.
$$P = \begin{cases} 2[1-\Phi(\lfloor z_0 \rfloor)] & \text{for a two} - \text{tailed test:} \quad H_o: \mu = \mu_0 \\ 1-\Phi(z_0) & \text{for a upper} - \text{tailed test:} \quad H_o: \mu = \mu_0 \\ \Phi(z_0) & \text{for a lower} - \text{tailed test:} \quad H_o: \mu = \mu_0 \\ \Phi(z_0) & \text{for a lower} - \text{tailed test:} \quad H_o: \mu = \mu_0 \\ \Phi(z_0) & \text{for a lower} - \text{tailed test:} \quad H_o: \mu = \mu_0 \\ \Phi(z_0) & \text{for a lower} - \text{tailed test:} \quad H_o: \mu = \mu_0 \\ \Phi(z_0) & \text{for a lower} - \text{tailed test:} \quad H_o: \mu = \mu_0 \\ \Phi(z_0) & \text{for a lower} - \text{tailed test:} \quad H_o: \mu = \mu_0 \\ \Phi(z_0) & \text{for a lower} - \text{tailed test:} \quad H_o: \mu = \mu_0 \\ \Phi(z_0) & \text{for a lower} - \text{tailed test:} \quad H_o: \mu = \mu_0 \\ \Phi(z_0) & \text{for a lower} - \text{tailed test:} \quad H_o: \mu = \mu_0 \\ \Phi(z_0) & \text{for a lower} - \text{tailed test:} \quad H_o: \mu = \mu_0 \\ \Phi(z_0) & \text{for a lower} - \text{tailed test:} \quad H_o: \mu = \mu_0 \\ \Phi(z_0) & \text{for a lower} - \text{tailed test:} \quad H_o: \mu = \mu_0 \\ \Phi(z_0) & \text{for a lower} - \text{tailed test:} \quad H_o: \mu = \mu_0 \\ \Phi(z_0) & \text{for a lower} - \text{tailed test:} \quad H_o: \mu = \mu_0 \\ \Phi(z_0) & \text{for a lower} - \text{tailed test:} \quad H_o: \mu = \mu_0 \\ \Phi(z_0) & \text{for a lower} - \text{tailed test:} \quad H_o: \mu = \mu_0 \\ \Phi(z_0) & \text{for a lower} - \text{tailed test:} \quad H_o: \mu = \mu_0 \\ \Phi(z_0) & \text{for a lower} - \text{tailed test:} \quad H_o: \mu = \mu_0 \\ \Phi(z_0) & \text{for a lower} - \text{tailed test:} \quad H_o: \mu = \mu_0 \\ \hline \hat{\beta}_0 = \hat{\beta}_0 + \hat{\beta}_1 x \\ \hline \hat{\beta}_1 = \hat{\beta}_1 \hat{\gamma}_1 + \hat{\gamma}_2 \hat{\gamma}_2 \hat{\gamma}_2 \\ \hline \hat{\beta}_1 = \hat{\gamma}_2 \hat$$

STANDARD STATISTICAL TABLES

1. Areas under the Normal Distribution

