Statistical Theory

Based on recorded lectures of prof. Ayala Cohen, Technion, 2012

Summarized by Ido Greenberg

Contents

Introduction

- **Probability** = [model + parameters → probability of data]
- (Parametric) **Statistics** = [model + data → parameters]
    - **Non-parametric statistics** – which does not assume any parametric model in advance – is out of the scope.

Descriptive statistics

- Mean, truncated average (ממוצע קטום), median, variance, std, range, quartiles (Q1, Q3), IRQ=Q3-Q1, quantiles.
    - החציון כגבול של ממוצעים קטומים עם קטימה השואפת ל-50%.
- **Asymmetry coefficient** ~ 3rd moment ~ $\sum(x_i - < x >)^3$ ~ which direction is the longer tail.
- Bar plot, hist (number of samples is proportional to space → width can be heterogeneous).
- **Box plot** – expresses size (median), dispersion (quartiles), possibly tails (min & max up to 2.5 IRQs), and exceptions (points beyond the 2.5 IRQs).
- **Q-Q plot** – compare two distributions by plotting quantile-vs-quantile.

- When comparing empirical dist' to theoretical one, it's conventional to plot x_i vs. $F\left(p = \frac{i}{n+1}\right)$, e.g. for 10 samples, x1 (after sort) represents F(1/11) and x10 represents F(10/11).
- Comparing $Y = N(\mu, \sigma^2)$ with $X = N(0,1)$ yields a line with intercept μ and incline σ, since $Y = \mu + \sigma X$. Thus when studying a possibly-normal empirical distribution, there's no need to estimate the parameters in advance – they can be QQ-plotted vs. standard normal dist'.

Inferential statistics

Introduction

- Notation conventions:
  - GREEK/greek = parameters Θ
  - ENGLISH = statistics X
  - $\widehat{\ }$ = estimators $\widehat{\Theta}$
  - english = values x
- Statistic = function of the known data (in particular doesn't directly depend on the parameters of the underlying dist')
- Estimation: statistic which estimates a parameter is an **estimator**, and its value for certain data is an **estimate**.
  - **Consistent** estimator – converges (in probability) to the parameter when n→inf.
    - Convergence in probability: $\forall \epsilon > 0: \ \Pr\left(|\hat{\theta}_n - \theta| > \epsilon\right) \to 0.$
  - **Unbiased** estimator – E[estimator]=parameter.

Estimation methods

- <u>**Moments estimation method**</u>: parameters can often be calculated **as function of the moments**, and the **moments can be consistently estimated** from data using simple estimators (means of powers).
  - For example: $\hat{\sigma} = \widehat{\sqrt{\mu_2 - \mu_1^2}} \coloneqq \sqrt{\widehat{\mu_2} - \widehat{\mu_1}^2}$ $\quad (\sigma^2 \coloneqq E((X-\mu)^2) = E(X^2) - E(X)^2)$
    - (estimator is chosen to be defined by moments' estimators)
  - Estimators defined by the moments method are **always consistent** (continuous function of consistent estimators…).
  - One should use as **low moments** as possible, since higher moments might have infinite expectations in certain cases.
- <u>Paradoxes in the moments method</u>:
  - $\widehat{\sigma^2} \coloneqq \widehat{\mu_2} - \widehat{\mu_1}^2$ is **consistent** (converges to σ^2) but **not unbiased**!
    - Since $\boldsymbol{E}\left(\widehat{\boldsymbol{\mu_1}}^{\mathbf{2}}\right) = E(\bar{X}^2) = Var(\bar{X}) + E(\bar{X})^2 = \frac{\sigma^2}{n} + \mu_1^2 \neq \boldsymbol{\mu_1^2}.$
    - Thus $\boldsymbol{E}\left(\widehat{\boldsymbol{\sigma^2}}\right) = \mu_2 - \frac{\sigma^2}{n} - \mu_1^2 = \sigma^2 - \frac{\sigma^2}{n} = \boldsymbol{\sigma^2}\left(\frac{\boldsymbol{n-1}}{\boldsymbol{n}}\right).$
    - Thus we choose $\widehat{\boldsymbol{\sigma^2}} \coloneqq \frac{\boldsymbol{n}}{\boldsymbol{n-1}}\left(\widehat{\boldsymbol{\mu_2}} - \widehat{\boldsymbol{\mu_1}}^{\mathbf{2}}\right)$ which is both consistent and unbiased (though not defined by the moments method).
      - Note: $\sqrt{\widehat{\boldsymbol{\sigma^2}}}$ **is <u>not</u> unbiased estimator for Standard Deviation!**

        - Bias explanation: Variance is measured using $\widehat{\mu_2}$ that **estimates** $\mu_2 = Var + \mu_1^2$, i.e. both the **dispersion of X (Var)** **and** its **squared bias (μ_1^2)**. To isolate the dispersion we **subtract the squared bias's estimator $\widehat{\mu_1}^2$**, but **due to the squaring it tends to overestimate** (since after squaring, 2→3 is larger error than 2→1), **thus the variance is underestimated** and requires the correction 1/n → 1/(n-1).
            - Moral: **expectation is sensitive to non-linear units-conversion such as squaring.**
    - **For $X \sim U(0,\theta)$, $\hat{\theta} := 2\widehat{\mu_1} = 2\bar{x}$** is consistent, even though it may be logically **impossible**!
        - E.g. for data (1,2,9) we have avg=4 thus $\hat{\theta} = 8$, though x3=9>8!
        - Note: Uniform distribution is often a simple example for anomalies.
- **Maximum likelihood**: $\hat{\theta} := argmax P(\{x\};\theta)$ – usually better than the moments method.

Properties of estimators

- **Bias** of estimator: $\boldsymbol{B_T(\theta) := E(T) - \theta}$
- **MSE** of Tn (estimator based on n samples): $\boldsymbol{MSE_{T_n}(\theta) := E\left[(T_n - \theta)^2\right]}$
    - Claim: [MSE(Tn)→0] ➔ [Tn→θ in probability] ➔ Tn is consistent
        - Proved directly by Chebyshev inequality.
    - Claim: $\boldsymbol{MSE_T(\theta) = Var(T) + B_T(\theta)^2}$ **= variance + bias**
        - Proved by adding +E(T)-E(T) within the definition of the MSE.
- Estimators of Uniform distribution [0, θ]:
    - $MSE_{2\bar{x}}(\theta) = bias^2 + Var = 0 + 4Var(\bar{x}) = \frac{\theta^2}{3n}$
    - $MSE_{\max(x_i)}(\theta) = bias^2 + Var = \cdots = \left(\frac{\theta}{n+1}\right)^2 + \cdots = O(\frac{1}{n^2})$ ➔ better
- Note:
    - **Consistency** of estimator is **preserved under continuous function** (as in the moments method).
    - **Unbiasedness** of estimator is **preserved under linear function**.
- **Risk** of estimator: $R := E\left(L\left(\hat{\theta},\theta\right)\right)$ for some Loss function L.

Sufficiency

- Sufficiency of estimator T_θ: $\boldsymbol{P(\{x_i\}|T_\theta)}$ **is independent of $\boldsymbol{\theta}$**.
    - Meaning: given T_θ, the dist' of the data is independent on θ
    - ➔ the raw data {x} doesn't provide additional information about θ
    - ➔ $T_\theta(\{x\})$ is sufficient to exploit all the information of {x} about θ.
        - **See also**: Fisher information, Observed information
    - E.g. in Bernoulli distribution, the rate of successes $\frac{\sum x_i}{n}$ is sufficient for estimation of p.
    - Note: statistic doesn't have to be scalar (e.g. $S := \{x\}$ is always sufficient for any θ…). **Minimal sufficient statistic** is a sufficient statistic of "minimal dimension" (formally – for any other sufficient T, it holds that $S = f(T)$).
- **Fisher-Neyman Factorization Theorem**: [S is sufficient wrt θ] iff [$f(\{x\};\theta) = h(\{x\}) \cdot \phi(S,\theta)$].
    - Example – normal distribution:

- If μ is known – then $\sum(x_i - \mu)$ is sufficient wrt σ: $f(\{\boldsymbol{x}\}, \boldsymbol{\mu}; \boldsymbol{\sigma}) = 1 \cdot \left(\frac{1}{\sqrt{2\pi\sigma^2}^n} e^{-\frac{\sum(x_i-\mu)^2}{2\sigma^2}}\right)$
- If σ is known – then $\bar{x}$ is sufficient wrt μ: $f(\{\boldsymbol{x}\}, \boldsymbol{\sigma}; \boldsymbol{\mu}) = \frac{1}{\sqrt{2\pi\sigma^2}^n} e^{-\frac{\sum(x_i-\bar{x})^2 + \sum(\bar{x}-\mu)^2}{2\sigma^2}} = \left(\frac{1}{\sqrt{2\pi\sigma^2}^n} e^{-\frac{\sum(x_i-\bar{x})^2}{2\sigma^2}}\right) \cdot \left(e^{-\frac{\sum(\bar{x}-\mu)^2}{2\sigma^2}}\right)$
- If both are unknown – then $\bar{x}$ and $\sum(x_i - \bar{x})$ together are a minimal sufficient statistic:
$$f(\{\boldsymbol{x}\}; \boldsymbol{\mu}, \boldsymbol{\sigma}) = 1 \cdot \left(\frac{1}{\sqrt{2\pi\sigma^2}^n} e^{-\frac{\sum(x_i-\bar{x})^2 + \sum(\bar{x}-\mu)^2}{2\sigma^2}}\right)$$

Sampling distributions

- What is the **required size of a sample set** intended to measure θ?
  - $\boldsymbol{min}\{\boldsymbol{n} \in \boldsymbol{N} \ \ | \ \ \boldsymbol{P}[|\boldsymbol{T_\theta} - \boldsymbol{\theta}| > \boldsymbol{d}] < \boldsymbol{\alpha}\}$ (for given d,α)
  - E.g. for μ in $N(\mu, \sigma^2)$: $P[|\bar{x} - \mu| > d] = 2\left(1 - \Phi\left(\frac{d\sqrt{n}}{\sigma}\right)\right)$ ➔ $\Phi\left(\frac{d\sqrt{n}}{\sigma}\right) > 1 - \frac{\alpha}{2}$

    ➔ $n > Z^2_{1-\frac{\alpha}{2}} \frac{\sigma^2}{d^2}$
  - This actually holds for any distribution, since $\bar{x} \to \mu$ by the **Central Limit Theorem**.

χ^2 distribution

- $\boldsymbol{\chi^2(n)}$: $\boldsymbol{f(y)} \coloneqq \frac{\boldsymbol{1}}{\Gamma\left(\frac{n}{2}\right)2^{\frac{n}{2}}} \boldsymbol{e^{-\frac{y}{2}} y^{\frac{n}{2}-1}}$ $(y \geq 0)$
  - n = "degrees of freedom"
  - n=2: $f(y) = \frac{1}{2} e^{-y/2}$ ➔ **generalization of exponential distribution**.
  - **Private case of Gamma distribution** with $\lambda = \frac{1}{2}, \alpha = \frac{n}{2}$:
$$\Gamma(\lambda, \alpha): \ f(y) \coloneqq \frac{\lambda^\alpha}{\Gamma(\alpha)} e^{-\lambda y} y^{\alpha-1}$$
  - $Z \sim N(0,1)$ ➔ $\boldsymbol{Z^2 \sim \chi^2(1)}$
  - $\boldsymbol{\sum Z_i^2 \sim \chi^2(n)}$ (sum of independent Gamma dists is calculated using moment-function?)
  - In general, for independent variables, $\boldsymbol{\chi^2(n_1) + \chi^2(n_2) = \chi^2(n_1 + n_2)}$
  - $E[y \sim \chi^2(n)] = n,$ $Var = 2n$
  - $\frac{\chi^2(n)}{n} \to 1$ (with probability) by Law of Large Numbers since $\chi^2(n) = \sum \chi^2(1)$
- Although $\sum\left(\frac{x_i-\mu}{\sigma}\right)^2 \sim \chi^2(n)$, without μ we "lose a degree of freedom", so $\sum\left(\frac{x_i-\bar{x}}{\sigma}\right)^2 \sim \chi^2(n-1)$
  - Equivalently for $\boldsymbol{s^2} \coloneqq \frac{\sum(x_i-\bar{x})^2}{\boldsymbol{n-1}}$, $\frac{\boldsymbol{(n-1)s^2}}{\boldsymbol{\sigma^2}} \sim \boldsymbol{\chi^2(n-1)}$
  - Proved through $\frac{(\bar{x}-\mu)^2}{\sigma^2/n} \sim \chi^2(1)$ and the fact that $\boldsymbol{\bar{x}, s}$ **are independent**

T-distribution and F-distribution

- $T \coloneqq \frac{Z}{\sqrt{\frac{w_k}{k}}} \sim t(k)$ $(Z \sim N(0,1), w_k \sim \chi^2(k))$
  - $T \to N(0,1)$ for k➔inf since $\frac{\chi^2}{n} \to 1$
  - $t_v^{1-\alpha} \coloneqq \arg(P(t(v) < X) = \alpha)$

  - $\frac{\bar{\boldsymbol{x}}-\boldsymbol{\mu}}{\boldsymbol{s}/\sqrt{\boldsymbol{n}}} \sim \boldsymbol{t}(\boldsymbol{n}-\mathbf{1})$
- $F_{m_1,m_2} \coloneqq \frac{\chi^2(m_1)/\sigma_1}{\chi^2(m_2)/\sigma_2}$
  - $F^{\alpha}_{m_1,m_2} = 1/F^{1-\alpha}_{m_2,m_1}$

Confidence interval

- **Pivotal quantity** (AKA **Pivot**): $f(\hat{\theta},\theta)$ whose distribution is the same for any θ.
- For $x_i \sim N(\mu,\sigma^2)$ with unknown params, $\bar{x}$ satisfies $P\left(\bar{\boldsymbol{x}} + \boldsymbol{t}^{\boldsymbol{n}-\mathbf{1}}_{\frac{\alpha}{2}} \frac{\boldsymbol{s}}{\sqrt{\boldsymbol{n}}} \le \boldsymbol{\mu} \le \bar{\boldsymbol{x}} + \boldsymbol{t}^{\boldsymbol{n}-\mathbf{1}}_{\mathbf{1}-\frac{\alpha}{2}} \frac{\boldsymbol{s}}{\sqrt{\boldsymbol{n}}} \mid \boldsymbol{\mu},\boldsymbol{s}\right) = \mathbf{1}-\boldsymbol{\alpha}$, independently of μ.
  - Note: that's the probability that $\bar{x}$ would be that close to μ (i.e. if we do many such experiments, we expect $\sim\alpha$ of the estimates to be that close to μ. **The probability that $\boldsymbol{\mu}$ lays within the confidence interval is defined only if a prior distribution is assumed on $\boldsymbol{\mu}$.**
  - $\frac{\boldsymbol{\mu}-\bar{\boldsymbol{x}}}{\boldsymbol{s}/\sqrt{\boldsymbol{n}}}$ **is a pivot** for μ with T-distribution.
  - $\left[\bar{x} + t^{n-1}_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}\ ,\ \bar{x} + t^{n-1}_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}\right]$ is a **2-sided *confidence-interval*** of μ with confidence $1-\alpha$.
  - The **symmetric 2-sided confidence-interval** is the **shortest interval** corresponding to a given confidence level – if the distribution of the estimator's distribution is symmetric around one maximum.
  - $\left[\bar{x}\ ,\ \bar{x} + t^{n-1}_{1-\alpha} \frac{s}{\sqrt{n}}\right]$ is a **1-sided** confidence-interval of μ with confidence $1-\alpha$.
- For two normally-distributed populations, one similarly has a confidence interval for the dispersion ratio $\frac{\sigma_2}{\sigma_1}$: $\left[\frac{s_2^2}{s_1^2} F^{\alpha/2}_{n_1-1,n_2-1}\ ,\ \frac{s_2^2}{s_1^2} F^{1-\alpha/2}_{n_1-1,n_2-1}\right]$
  - Relevant to measure ratio between diversions of two populations – e.g. men & women salaries, or errors of two different measurement devices.
  - Note: in this case the symmetric interval is not the shortest (since F is a-symmetric), but is just the quickest to calculate.
- For two **independent** normal variables:
  - $\bar{\boldsymbol{x}} - \bar{\boldsymbol{y}} \sim N\left(\mu_1-\mu_2, \frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right)$
  - If $\sigma_1 = \sigma_2$ then $\frac{(\bar{x}-\bar{y})-(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim N(0,1)$. One can prove that replacing the (typically unknown) σ with $\hat{\sigma} \coloneqq s$ (weighted average of s_1 and s_2, which is $\chi^2(n_1+n_2-2)$) yields T-distribution with n_1+n_2-2 DoF.
  - The confidence interval: $\mu_1-\mu_2 \in \bar{x}-\bar{y} \pm s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\, t^{1-\frac{\alpha}{2}}_{n_1+n_2-2}$
- For two **dependent** normal variables:
  - If x,y are set to have Cor=ρ>0, then the confidence interval can be smaller.
    - This is called ***Blocking*** in experimental statistics, named after choosing similar blocks for agricultural experiments.
  - $\sigma_D^2 \coloneqq Var(x_i - y_i) = \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2$
    - $\rightarrow x - y \sim N(\mu_1-\mu_2\ ,\ \sigma_D^2)$

  - $\widehat{\sigma_D^2} \coloneqq s_D$ (s for the samples $D_i \coloneqq x_i - y_i$)
  - Now the confidence interval can be calculated as in the normal-variable case.

Hypothesis tests

- **Accepting** the **Null-Hypothesis** H0 ←→ Data {xi} are reasonably consistent with H0 ←→ $\{x_i\} \notin R$ where $P(!\,R|H_0) = 1 - \alpha$.
- **Rejecting** H0 in favor of an **Alternative-Hypothesis** H1 ←→ $\{x_i\} \in R$.
  - It is a-priory assumed that either H0 is true or H1 is true.
- Simple hypothesis = specific distribution; composite hypothesis = family of distributions.
  - P=1/2 vs. P>1/2 is one-sided test of simple hypothesis Vs. composite hypothesis.
- Errors:
  - **Type 1** ($P = \alpha$) = false rejection ("radical") **Significance**=$P(R|H_0) = \alpha$
  - **Type 2** ($P = \beta$) = false acceptance ("conservative") **Power** = $P(R|H_1) = 1 - \beta$
  - **Significant** (small α) = being "fair" with H0 – not rejecting in vain.
  - **Powerful** (small β) = being "open" to rejecting H0 in favor of H1.
- Rule of decision (R) = set of test results for which H0 will be rejected.
  - R1 is *better* than R2 if $\alpha_1 \le \alpha_2$ & $\beta_1 \le \beta_2$.
  - R is *admissible* if no R' is better.
- Setting R given α:
  - If the range of the data samples is continuous or dense – then R can be defined in terms of thresholds on the data.
  - **If the range of the data is discrete** – with some admissible rules "too" significant (smaller α, hence unnecessarily larger β) and some not enough significant (larger α) – then the exact threshold α can be achieved by a ***mixed rule*** that randomly chooses one of two ***pure rules*** (with corresponding probabilities).
- **Neyman-Pearson Lemma**:
  - ***Likelihood ratio***: $\lambda(x) \coloneqq P(x|H_1)/P(x|H_0)$
  - NP Lemma: for simple-vs.-simple hypothesis test with **Significance**$\le \boldsymbol{\alpha}$**, the maximal power is achieved by** $\phi(x) \coloneqq P(reject) \coloneqq \begin{cases} 1 & if\ \lambda(x) > k_\alpha \\ \Gamma_\alpha & if\ \lambda(x) = k_\alpha \\ 0 & if\ \lambda(x) < k_\alpha \end{cases}$ for a certain k_α – i.e. by **determining a threshold depending on the required $\boldsymbol{\alpha}$**, and in the discrete case – possibly having random choice in the threshold itself.
- ***P-value***:
  - $p = \text{argmin}_\alpha(x \in R_\alpha) =$
    $\text{how significant (conservative, "fair") can we be while still rejecting } H_0 =$
    $\text{how conservative (small } \alpha\text{) we need to be to yet accept } H_0 =$
    $P(\text{having results "}\textit{as extreme as}\text{" } x \mid H_0)$
  - P-value deals only with type-I error – it's independent of H1.
- Composite hypotheses test:
  - E.g. $H_0 \coloneqq \mu \le \mu_0$ vs. $H_1 \coloneqq \mu > \mu_0$, or $H_0 \coloneqq \mu = \mu_0$ vs. $H_1 \coloneqq \mu \ne \mu_0$.
  - In general: $\boldsymbol{H_0 \coloneqq (\theta \in \omega)}$**, and** $\boldsymbol{\alpha \coloneqq \sup_{\theta \in \omega} P_\theta(R)}$.
  - Note: a **confidence interval of confidence** $\boldsymbol{1 - \alpha}$ around $\bar{x}$ contains all the values μ_0 of μ for which the data **{xi} do not reject the hypothesis** $\boldsymbol{\mu = \mu_0}$ **with significance** $\boldsymbol{\alpha}$.

- *Generalized likelihood ratio*: $\Lambda(x) := \frac{\sup_{\theta\in\omega} f_\theta(x)}{\sup_{\theta\in\Omega} f_\theta(x)}$ ($\theta \in \omega$ is H0, $\Omega = dom(\theta)$)
    - Generalized likelihood ratio test: $\Lambda(x) < k_\alpha$.
    - Likelihood ratio for composite HT: $\lambda(x) := \frac{\sup_{\theta\in\omega^c} f_\theta(x)}{\sup_{\theta\in\omega} f_\theta(x)}$
        - These *sups* are achieved by ML estimates for θ.
    - For normal distribution with unknown σ and $H_0: \mu = \mu_0$, we have:
        - $\sup_{\theta\in\omega} f_\theta(x)$ is achieved by $\hat{\sigma} = s := \frac{1}{n}\sum(x_i - \boldsymbol{\mu_0})^2$
        - $\sup_{\theta\in\Omega} f_\theta(x)$ is achieved by $\hat{\sigma} = s := \frac{1}{n}\sum(x_i - \boldsymbol{\bar{x}})^2$
- Rejection rule with significance α for $H_0 := \mu = \mu_0$ vs. $H_1 := \mu \neq \mu_0$: $\left|\frac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}}\right| > t_{n-1}^{1-\frac{\alpha}{2}}$

Fit tests

- **Theorem**: the generalized likelihood ratio asymptotically satisfies $\boldsymbol{\Lambda^* := -2\ln(\Lambda) \sim \chi^2(n)}$, where $n = \dim(\Omega) - \dim(\omega)$.
    - Difference of dimensions **n** is actually the **number of constraints in the model corresponding to $\boldsymbol{\omega}$**.
    - E.g. if we claim that $\mu = \mu_0$ & $\sigma = \sigma_0$ then $n = \dim(\Omega) - \dim(\omega) = 2 - 0 = 2$.
- ***Fit-test***: given data and a possible **discrete model**, one can **calculate the likelihood of the data and the maximum likelihood**, and **test the hypothesis that the data is generated in accordance with the model**.
    - A continuous model can be tested by approximating it to discrete values (as in histogram).
    - **Example**: dice with H0 := fair dice, and N rolls with xi:=#(rolls with result i):
        - The maximum likelihood is achieved for $P_i^{ML} := x_i/N$.
        - The statistic is $\Lambda^* = -2\ln(\Lambda) = -2\sum_{i=1}^{6} x_i \ln\left(P_i^0/P_i^{ML}\right)$.
        - The distribution under H0 is $\chi^2(5-0) = \chi^2(5)$, from which one can get p-value.
    - **In general for simple hypothesis $\boldsymbol{H_0 = \{p_i^0\}_{i=1}^n}$ on parameters space with $\boldsymbol{\dim(\Omega) = n}$, and N samples, we have the asymptotic distribution:**

$$\boldsymbol{\Lambda^* = -2\sum_{i=1}^{n} x_i \ln\left(\frac{p_i^0 N}{x_i}\right) \sim \chi^2(n)}$$

- **Approximated $\boldsymbol{\chi^2}$-test**:
    - $X_p^2 := \sum\left(\frac{(O_i-E_i)^2}{E_i}\right) \to \Lambda^*$ (they have the same asymptotic distribution)
        - Ei = expected i'th value under H0 = $N \cdot p_i^0$
        - Oi = observed i'th value = x_i
    - Proved directly by ln(1+x) ~ x-0.5*x^2.
    - Poor approximation for any $E_i < 5$. This can be avoided by uniting values-categories.

Independence tests

- Independence test between X1,X2 **can be seen as fit test** to the hypothesis of independence.
- Formalization:

- Values-categories: $\{ij\}_{i=1:K_1,j=1:K_2}$ (assuming that X1,X2 are K1,K2-discrete)
- ML: $P_{ij}^{ML} = x_{ij}$ (ML of all pairs) or $O_{ij} = x_{ij}$
- H0: $P_{ij}^{0} = \frac{x_{i*}}{N} \cdot \frac{x_{*j}}{N}$ (ML of i times ML of j) or $E_{ij} = NP_{ij}^{0}$
- DoF: $(K_1K_2 - 1) - \left((K_1 - 1) + (K_2 - 1)\right) = (K_1 - 1)(K_2 - 1)$

* Note: testing whether a parameter is identical over 2 populations can be done now using independence test rather than F-test of the ratio.

Linear regression

* Predicting an ***dependent*** variable Y using ***explanatory/independent*** variable X.
* ***Regression function***: $g(x) \coloneqq E[Y|X = x]$ ("value-per-quanta")
* Linear regression model:
  - $\epsilon_i \coloneqq (y_i - \alpha - \beta x_i) \sim N(0, \sigma^2)$ (*homoscedasticity* = σ is independent of x)
  - Cov(ϵ_i, ϵ_j)=0 for i≠j (in particular not time series)
  - Goal: estimate $\boldsymbol{a} \coloneqq \boldsymbol{\hat{\alpha}}$, $\boldsymbol{b} \coloneqq \boldsymbol{\hat{\beta}}$, $\boldsymbol{s} \coloneqq \boldsymbol{\hat{\sigma}}$
  - Notation: $e_i \coloneqq \hat{\epsilon_i} \coloneqq y_i - \hat{y_i} = y_i - a - bx_i$
* Note: **<u>linearity</u> and <u>independence</u> are strong and mostly unrealistic assumptions**.
* **Least squares**:
  - $a, b \coloneqq argmin\left(\sum e_i^2\right)$
  - Solution (derive and compare to 0):
    - $b = \cdots = \frac{\sum(y_i - \bar{y})x_i}{\sum(x_i - \bar{x})x_i} = \cdots = \sum \frac{(x_i - \bar{x})y_i}{\sum(x_i - \bar{x})^2} = \sum w_i y_i$ $\left(\sum w_i = 0, \sum w_i x_i = 1\right)$
    - $a = \bar{y} - b\bar{x}$
  - Note: $S_{xy} := \sum\left((x_i - \bar{x})(y_i - \bar{y})\right) \rightarrow b = S_{xy}/S_{xx}$ (inconvenient formulation…)
    - $S_{xx} = \sum(x_i - \bar{x})^2$
  - Note: the regression line always passes through $(\bar{x}, \bar{y})$.
* LS coefficients estimation statistics:
  - $E(b) = \sum w_i E(y_i) = \alpha \sum w_i + \beta \sum w_i x_i = \beta$ (unbiased estimator)
  - $Var(b) = \sum w_i^2 Var(y_i) = \sigma^2 \sum w_i^2 = \sigma^2 / S_{xx}$
  - ➔ $\boldsymbol{b \sim N\left(\beta, \sigma^2 / S_{xx}\right)}$
  - Note: **b is most accurate** when Sxx is maximal, which is achieved **by choosing the xi to be as far as possible in the edges of dom(x)**. This is indeed the way to have accurate estimation of a line, but it prevent us from judging whether it's indeed a line (i.e. whether the linear model is reasonable).
  - Similarly:
    - $\boldsymbol{a \sim N\left(\alpha, \sigma^2 \frac{\sum x^2}{NS_{xx}}\right)}$
    - $s^2 = \frac{\sum e_i^2}{N-2}$, $\frac{(N-2)s^2}{\sigma^2} \sim \chi^2(N-2)$
  - $\frac{\hat{\beta} - \beta}{\hat{\sigma_\beta}} \sim t_{N-2}$
    - In particular for $\boldsymbol{H_0: \beta = 0}$, one has: $\boldsymbol{\frac{\hat{\beta}}{\hat{\sigma}}\sqrt{S_{xx}} \sim t_{N-2}}$
    - This derives a **regression test** with the α-confidence interval:

$$\hat{\beta} - t_{1-\frac{\alpha}{2}}^{N-2}\widehat{\sigma_\beta} \le \beta \le \hat{\beta} + t_{1-\frac{\alpha}{2}}^{N-2}\widehat{\sigma_\beta}$$

    - Equivalently, $\frac{\hat{\beta}^2 S_{xx}}{\hat{\sigma}^2} \sim F_{1,N-2}$.
- Prediction: $\hat{\boldsymbol{y}} = a + bx \sim \boldsymbol{N}\left(\boldsymbol{y}, \boldsymbol{\sigma}^2\left(\boldsymbol{1} + \frac{\boldsymbol{1}}{\boldsymbol{N}} + \frac{(\boldsymbol{x}-\bar{\boldsymbol{x}})^2}{\boldsymbol{S_{xx}}}\right)\right)$ (under the linear regression model)
  - The error of $\hat{y}$ consists of 3 terms:
    - Inherent noise in the model (1)
    - Error in the estimate of α (1/N) – smaller for larger N
    - Error in the estimate of β ($\frac{(x-\bar{x})^2}{S_{xx}}$) – smaller for either larger N or x's closer to $\bar{x}$
  - Note: unlike prediction ("what will be y for a certain x0?") – which is affected directly by the noise σ – estimation of the expectation $E[y|x_0]$ ("what is the average y over the x=x0 population?") is affected by the noise only through the errors in the parameters estimates, thus such estimation will use the variance $\sigma^2\left(\frac{1}{N} + \frac{(x-\bar{x})^2}{S_{xx}}\right)$ (without the "1").
    - In other words, significance interval for prediction is wider than significance interval for parameter estimation.
- Analysis of Variance (***ANOVA***):
  - $\sum e_i^2 = S_{yy} + \hat{\beta}^2 S_{xx} - 2\hat{\beta} S_{xy} = S_{yy} - \hat{\beta}^2 S_{xx}$
  - Equivalently, $\boldsymbol{S_{yy} = \hat{\beta}^2 S_{xx} + \sum e_i^2}$, i.e. the **variance of Y (n-1 DFs)** is partially **explained by X (regression variance, 1 DF)**, and partially **unexplained (residuals variance, n-2 DFs)**.
  - **The part of Y which is explained by X:** $\boldsymbol{R^2 := \frac{\hat{\beta}^2 S_{xx}}{S_{yy}} = \frac{S_{xy}^2}{S_{xx} S_{yy}}}$
    - Note: the last formulation is symmetric between X & Y.
  - $F := \frac{\hat{\beta}^2 S_{xx}}{\hat{\sigma}^2} = (N-2)\frac{R^2}{1-R^2} \sim F_{1,N-2}$ is a statistic useful for regression F-test.
- Two regression lines – Y/X vs. X/Y:
  - $\beta_{y/x} = S_{xy}/S_{xx}$ whereas $\beta_{x/y} = S_{xy}/S_{yy}$.
  - $\beta_{y/x} \cdot \beta_{x/y} = R^2$ = 1 iff the regression lines are identical.
  - $\beta_{y/x}/\beta_{x/y} = S_{yy}/S_{xx}$ = scales ratio
- Correlation coefficient: $R := \frac{\frac{S_{xy}}{N}}{\sqrt{\frac{S_{xx}}{N} \cdot \frac{S_{yy}}{N}}} \to \frac{Cov(x,y)}{Var(x)Var(y)} = \rho$
  - Note: R is a consistent but biased estimator of ρ.
  - Note: $\boldsymbol{\rho = 0 \ iff \ \beta = 0}$.
- **Scaling:**
  - $\boldsymbol{\hat{\beta}_{vy/ux} = \frac{v}{u}\hat{\beta}_{y/x}}$
  - $\boldsymbol{\hat{\alpha}_{vy/ux} = v\hat{\alpha}_{y/x}}$
  - $\boldsymbol{R_{vy/ux} = sign(uv) R_{y/x}}$
- Multi-regression: regression with multiple variables.
  - Note: a model containing non-linear powers of a variable can be linearized by referring to X^p as a new variable with linear relation to Y.
- Terminology:
  - The linear regression model is **asymmetric – all the errors are associated with Y** (since we minimize vertical errors rather than geometrical distance of the samples from the line).

        - That's why the two regression lines differ unless $R^2 = 1$.
    - Since a **linear regression model explains (through X) only part of the variance of Y**, then **the dispersion of $\hat{y}$ around the mean $\bar{y}$ will always be smaller than the dispersion of the true y** – thus the model suggests a **regression of the phenomenon of y towards its mean**.
    - Algebraically: $\boldsymbol{S_{\hat{y}\hat{y}}} = \hat{\beta}^2 S_{xx} = \frac{S_{xy}^2}{S_{xx}} \le \frac{S_{xx}S_{yy}}{S_{xx}} = \boldsymbol{S_{yy}}$.
        - $\boldsymbol{R^2 = 1} \quad \rightarrow \quad \hat{y} = y \,, S_{xy}^2 = S_{xx}S_{yy} \quad \rightarrow \quad S_{\hat{y}\hat{y}} = S_{yy}$ (full reconstruction)
        - $\boldsymbol{R^2 = 0} \quad \rightarrow \quad \beta = 0 \quad \rightarrow \quad \hat{y} \equiv \bar{y} \,, S_{\hat{y}\hat{y}} = 0$ (full regression to mean)
    - Historically

Degrees of Freedom

- The number of *degrees of freedom* of a dynamic system is the number of independent ways by which its input can move without violating any constraint imposed on it.
    - A dynamic system is not a statistical term, though, and the conversion to statistics is unclear.
- A statistic is a function of data: $\boldsymbol{S = f(\{x_i\}_{i=1}^{n})}$
- A statistic is often defined as an estimator of an unknown parameter.
- **Degrees of freedom of an estimate is the number of independent pieces of information that went into calculating the estimate.**
    - Which is of course an ambiguous definition, e.g. variance can be seen as $\sum(x_i - \bar{x})^2$ (n), $\sum t_i^2 + (\sum t_i)^2$ (n-1) or just s (1) – all are calculations of independent elements…
- Many statistics are **commutative functions** of the data (i.e. independent of the order, e.g. mean & variance). In addition, it is often assumed that the data samples are **i.i.d**.
- Under such assumptions, the **distribution of S depends** on the **distribution of each sample** and on the **number of data samples (n)**.
- For additive statistic (e.g. $S = \sum x_i$ or $S = \sum x_i^2$), the distribution is typically wider as n is larger. It is said that the statistic has n degrees of freedom to vary and add to the statistic.
- There are also statistics which are additive function of some variation of the input, e.g. $S = \sum(x_i - \bar{x})^2$.
    - Note: the input consists of n variables, but only n-1 of the additive terms are independent – the last one is determined deterministically by the sum of the others. It indeed turns out to narrow the distribution accordingly - $\sum_1^n(x_i - \bar{x})^2$ has the same distribution as $\sum_1^{n-1}(x_i - \mu)^2$.
    - However, this intuition is hard to formulate, and until now any case I saw of statistic whose distribution has certain "DFs", required a dedicated formulation and mathematical proof.
    - Indeed, statistical DFs are most commonly associated with the distributions t, F, χ^2.
- Note: statistical DFs are quite opposite to the intuition of modeling, in which the parameters are degrees of freedom of the model, and the data samples are the (weak) constraints. Here **the data has DFs that "help" it to get complex, while we use models to constraint its variety**. The residuals of the model always have less DFs to deviate from the model.
    - Actually, one separates model DFs from residuals DFs, and the sum is the data DFs. So there's kind of symmetric perspective of the DFs.
    - In linear regression (with intercept), DF(model)=2 and DF(residuals)=n-2.

- In **generalized or regularized linear models**, ***effective DFs*** can be defined using the hat-matrix (defined by $\hat{\boldsymbol{y}} = \boldsymbol{Hy}$), as **DF(model)=tr(H)**. It can be seen as **"how much the (Y) data can potentially affect the model predictions"** (sum of influences of samples).
- For example in ridge regression $\hat{\boldsymbol{\beta}} \coloneqq \left(\boldsymbol{X^T X} + \boldsymbol{\lambda I}\right)^{-1} \boldsymbol{X^T y}$, thus DF=$tr(X(X^T X + \lambda I)^{-1} X^T)$ which **deviates down from $\boldsymbol{m}$ as $\boldsymbol{\lambda}$ gets farther from 0** (m=#variables; it's 1 for single-input regression without intercept).
- Bonus: **regularization as a solution to ML problem**:

$$\boldsymbol{P(Y, b|X)} = P(b|X)P(Y|b, X) \sim e^{-||Y-Xb||^{p_1}/2\sigma^{p_1}} e^{-||b||^{p_2}/2\tau^{p_2}}$$

$$argmaxP(Y, b|X) = argmin(-logP) = argmin\left(||Y - Xb||^{p_1} + \frac{\sigma^{p_1}}{\tau^{p_2}}||b||^{p_2}\right)$$

$$= argmin_{Y,b}\left(||Y - bX||^{p_1} + \lambda||b||^{p_2}\right) \qquad \left(\lambda = \frac{\sigma^{p_1}}{\tau^{p_2}} = \frac{noise}{\beta s\text{-}power}\right)$$

- See also: DFs vs. complexity.