Lecture 04: Classical Inferential Statistics II: Significance Tests

4. Significance Tests
4.1 General Four-Step Procedure

1. Formulate a **null hypothesis H_0** such that their rejection gives insight, e.g. $\beta_j = \beta_{j0}$ (point hypothesis) or $\beta_j \leq \beta_0$ (interval hypothesis): Notice: *One cannot confirm* H_0
2. Select a **test function** or **statistics T**
   - whose distribution is known provided the parameters are at the **margin H_0^* of the null hypothesis** (of course, $H_0^* = H_0$ for a point null hypothesis)

     What if the estimator has a known distribution but the variance is unknown?
     Test function in units of the estimated standard deviation
   - which has distinct **rejection regions $R(\alpha)$** which are reached rarely (with a probability $\leq \alpha$) if H_0 but more often if $H_1 = \overline{H_0}$
3. Evaluate a realisation t_{data} of T from the data
4. Check if $t_{\text{data}} \in R(\alpha)$. If yes, H_0 can be rejected at an error probability or **significance level α**. Otherwise, *nothing can be said* (mask example with H_0: "mask useless").

4a Alternatively, calculate the ***p*-value** as the minimum α at which H_0 can be rejected.

4.1.1 Step 1: Choosing H_0: Type I and II errors

	H_0 not rejected	H_0 rejected
H_0 is true	✓	Type I error
H_0 is not true	Type II error	✓

- A significance test reduces reality to a "binary in-binary out" setting. There are two combinations corresponding to a correct test result
- We can control the **type I or α-error** probability $P(H_0$ rejected$|H_0) \le \alpha$ in **significance tests**
- Since the **type II or β-error** probability $P(H_0$ not rejected$|\overline{H_0})$ is unknown, the more serious error type should be the α error
- Fundamental problem: I want $P(H_0|$rejected$)$ and $P(H_0|\overline{\text{rejected}})$ while I get control over $P(\text{rejected}|H_0) \le P(\text{rejected}|H_0^*) \Rightarrow$ **Bayesian statistics**

4.1.2 Steps 2 and 3: Test statistics I

- (i) Testing parameters such as H_0: $\beta_j = \beta_{j0}$ or $\beta_j \geq \beta_{j0}$ or $\beta_j \leq \beta_{j0}$: The test function is the estimated deviation from H_0^* in units of the estimated error standard deviation. It is **student-t** distributed with #dataPoints- #parameters **degrees of freedom (df)**:

$$T = \frac{\hat{\beta}_j - \beta_{j0}}{\sqrt{\hat{V}_{jj}}} \sim T(n-1-J)$$

- (ii) Testing functions of parameters such as H_0: $\beta_1/\beta_2 = 2$, ≤ 2 or ≥ 2: Transform into a linear combination. Then, the normalized estimated deviation is student-t distributed under H_0^*. Here, at H_0^*, the linear combination is $b = \beta_1 - 2\beta_2 = 0$:

$$\begin{aligned} \hat{b} &= \hat{\beta}_1 - 2\hat{\beta}_2, \\ \hat{V}(\hat{b}) &= \hat{V}_{11} + 4\hat{V}_{22} - 4\hat{V}_{12}, \\ T &= \frac{\hat{b}}{\sqrt{\hat{V}(\hat{b})}} \sim T(n-1-J) \end{aligned}$$

Test statistics II

- (iii) Testing the correlation coefficient in an xy scatter plot:

$$\hat{\rho} = \frac{s_{xy}}{s_x s_y}, \quad H_0: \rho = 0, \quad T = \frac{\hat{\rho}}{\sqrt{1-\hat{\rho}^2}}\sqrt{n-2} \sim T(n-2)$$

Derivation: $\rho = 0$ if, and only if, in a simple linear regression $y = \beta_0 + \beta_1 x + \epsilon$, the slope parameter $\beta_1 = 0$, so test for $\beta_1 = 0$: Under H_0, the test statistics

$$T = \hat{\beta}_1 / \sqrt{\hat{V}_{11}} = \frac{s_{xy}}{\hat{\sigma}\ s_x}\sqrt{n} \sim T(n-2)$$

Now insert $\hat{\sigma}$ which can, in the simple-regression case, be explicitely calculated: $\hat{\sigma}^2 = n(s_y^2 - s_{xy}^2/s_x^2)/(n-2)$

- (iv) Test for the residual variance, H_0: $\sigma^2 = \sigma_0^2$, $\sigma^2 \geq \sigma_0^2$, and $\sigma^2 \leq \sigma_0^2$:

$$T = \frac{\hat{\sigma}^2}{\sigma_0^2}\ (n-1-J) \sim \chi^2(n-1-J)$$

The one-parameter **chi-squared distribution with m degrees of freedom** $\chi^2(m) = \sum_{i=1}^m Z_i^2$ is the sum of squares of i.i.d. Gaussians. *Its density is not symmetric, so we need to calculate both the α and $1-\alpha$ quantiles*

Test statistics III

- (v) Tests of simultaneous point null hypotheses, e.g., H_0: $(\beta_1 = 0)$ AND $(\beta_2 = 2)$ using the **Fisher-F test**:

$$T = \frac{(S_0 - S)/(M - M_0)}{S/(n - M)} \sim F(M - M_0, n - M)$$

  - S: SSE of the estimated full model with $M = J + 1$ parameters
  - S_0: SSE of the estimated restrained model under H_0 with M_0 free parameters

- The **Fisher-F** distribution is essentially the ratio of two independent χ^2 distributed random variables,

$$F(n, d) = \frac{\chi_n^2/n}{\chi_d^2/d},$$

with n numerator and d denominator degrees of freedom

? Argue that always $S_0 \geq S$

Equivalence of the F and T-tests for one parameter

With $M - M_0 = 1$, the F-test is equivalent to a parameter test for the parameter j in question:

- Parameter test: $T = \frac{\hat{\beta}_j - \beta_{j0}}{\sqrt{\hat{V}(\hat{\beta}_j)}} \sim T(n-1-J)$
- F-test: $T = (n-J-1)\frac{S_0 - S}{S} \sim F(1, n-1-J)$

? Regarding the rhs., show following general relation between the student-t and the F(1,d) distributions: $F \sim F(1,d)$ and $T \sim T(d) \Rightarrow F = T^2$

! By definition, Fisher's F is a ratio of χ^2 distributions. Furthermore, squares of standardnormal random variables Z are χ^2_1 distributed:

$$F(1,d) = \chi^2_1/(\chi^2_d/d) = Z^2/(\chi^2_d/d)$$

where $Z \sim N(0,1)$ and χ^2_d and Z are independent from each other. The definition of the student-t distribution is $T(d) = Z/\sqrt{\chi^2_d/d}$, so $F(1,d) = T^2_d$.

- One can show (difficult!) that following is exactly valid for the lhs.:

$$(n-J-1)\frac{S_0 - S}{S} = \frac{(\hat{\beta}_j - \beta_{j0})^2}{\hat{V}(\hat{\beta}_j)} = \frac{(\hat{\beta}_j - \beta_{j0})^2}{\hat{V}_{jj}}$$

where S_0 is the (minimum) SSE for the calibrated restrained model

4.1.3 Step 4: Decision

- The decision is based on the *rejection region*:

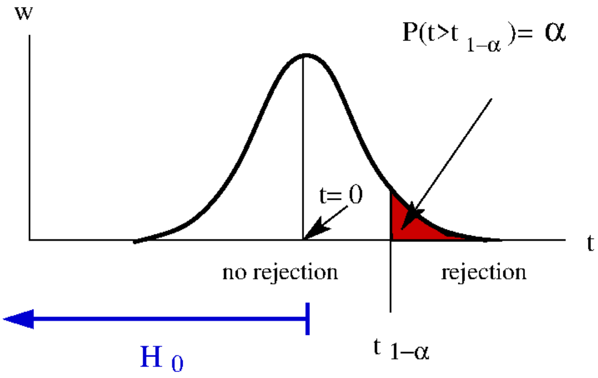
> The **rejection region $R^{(H_0)}(\alpha)$** contains the fraction α of all realisations t of the test statistics T which, under H_0^*, are most distant from H_0

- Decision:

> H_0 is rejected at significance level α if $t_{\text{data}} \in R^{(H_0)}(\alpha)$

- A good test statistics allows for a clear definition of what is meant by "distance to H_0" and brings, for a given α, the boundary of the rejection region as close to H_0^* as possible
- In contrast to T and the realisation t_{data} which only depends on H_0^* and therefore is the same for point and interval hypotheses of the same kind, the rejection region is different for the different comparison operators $=$, $\geq$, $\leq$

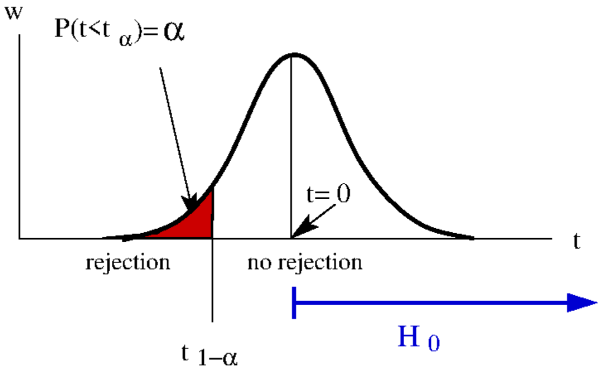
1. Rejection region for H_0: "$<$" or "$\leq$" (interval hypothesis)



- H_0 is rejected on the level α if

$$t_{\text{data}} > t_{1-\alpha}$$

2. Rejection region for H_0: "$>$" or "$\geq$" (interval hypothesis)

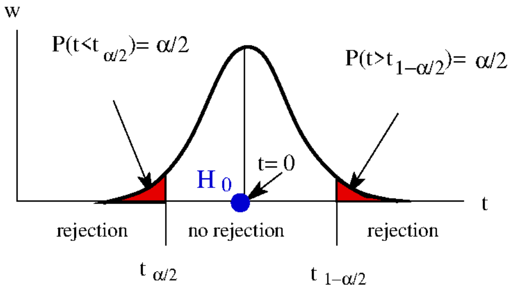


- H_0 is rejected on the level α if

$$t_{\text{data}} < t_\alpha = -t_{1-\alpha}$$

- The equality sign is only valid for symmetric test statistics

3. Rejection region for H_0: "=" (point hypothesis)



- For symmetric test statistics, H_0 is rejected on the level α if

$$|t_{\text{data}}| > t_{1-\alpha/2}$$

- If the distribution is not symmetric (as the χ^2 distribution for the variance test), the definition of what is "most distant" is not unique. For simplicity, one assumes equal statistical weights to both sides:

$$\text{rejected} \quad \Leftrightarrow (t_{\text{data}} < t_{\alpha/2}) \cup (t_{\text{data}} > t_{1-\alpha/2})$$

Example: modeling the demand for hotel rooms

The already well-known example for $y(\boldsymbol{x})$: hotel room occupancy [%]

$$y = \beta_0 x_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

where $x_0 = 1$, x_1: proxy for quality [# stars]; x_2: price [€/night],

$$\hat{\beta}_0 = 25.5, \quad \hat{\beta}_1 = 38.2, \quad \hat{\beta}_2 = -0.952$$

and

$$\hat{\mathbf{V}} = \begin{pmatrix} 28.0 & -6.40 & -0.119 \\ -6.40 & 26.0 & -0.941 \\ -0.119 & -0.941 & 0.0397 \end{pmatrix}$$

- ? Formulate and test the null hypothesis at $\alpha = 5\,\%$ that the stars do not matter
- ! $H_{01} : \beta_1 = 0$, point t-test with $T = \hat{\beta}_1 / \sqrt{\hat{V}_{11}} \sim T(12-3)$, i.e. df=9 degrees of freedom, $t_{\text{data}} = 7.49$, $t^{(9)}_{0.975} = 2.26 < |t_{\text{data}}| \Rightarrow H_0$ rejected, stars matter
- ? Do people favour more stars (at $\alpha = 5\,\%$)?
- ! $H_{02} : \beta_1 <= 0$ (use as H_0 what you want to reject!), interval test with same T and t_{data} as above, $t^{(9)}_{0.95} = 1.83 < t_{\text{data}} \Rightarrow H_{02}$ rejected, more stars are better

Example: modeling the demand for hotel rooms (ctned)

? Does each € more per night decrease the occupancy by at most 1 %?

! $H_{03}: \beta_2 < -1$ (H_{03} is the complement event!),
$t_{\text{data}} = (\hat{\beta}_2 + 1)/\sqrt{\hat{V}_{22}} = 0.24 \overset{!}{>} t^{(9)}_{0.95} = 1.83 \Rightarrow H_{03}$ not rejected
$\Rightarrow$ the hotel manager might risk losing more than one percent point of customers

? Is it worth renovating my hotel thereby gaining one star so that I can ask for 30 € more per night without losing guests?

! Again, define the complement event as $H_{04}: \beta_1 \le -30\beta_2$ or $\gamma = \beta_1 + 30\beta_2 \le 0$

$$\begin{aligned} \hat{\gamma} &= \hat{\beta}_1 + 30\hat{\beta}_2 = 9.63, \\ \hat{V}(\hat{\gamma}) &= \hat{V}_{11} + 900\hat{V}_{22} + 2 * 1 * 30\hat{V}_{12} = 5.27 \end{aligned}$$

So, $t_{\text{data}} = \hat{\gamma}/\sqrt{\hat{V}(\hat{\gamma})} = 4.20 > t^{(9)}_{0.95} = 1.83 \Rightarrow H_{04}$ rejected at 5 % $\Rightarrow$ the risk of losing customers is less than 5 %

? Can it be simultaneously true that $\beta_1 = 30$ and $\beta_2 = -1$?

! Full model: $\hat{\boldsymbol{\beta}} = (25.5, 38.2, -0.952)'$, $S(\hat{\boldsymbol{\beta}}) = 498.2$;
Reduced model with fixed $\beta_1 = 30$, $\beta_2 = 1$ leading to $\hat{\beta}_0 = 49.0$:
$\hat{\boldsymbol{\beta}}_r = (49.0, 30, -1)'$, $S_0 = S(\hat{\boldsymbol{\beta}}_r) = 1808$; $M - M_0 = 2\,\text{df}$, $n - M = 9\,\text{df}$,
$T \sim F(2, 9)$, $t_{\text{data}} = 9/2\ (S_0 - S)/S = 11.8 > f^{(2.9)}_{0.95} = 4.26 \Rightarrow H_0$ rejected

4.1.4 The p-value

- Obviously, it is not very efficient to test H_0 for a fixed significance level α (one does not know *how significant* the result really is)
- Instead, one would like to know the *minimum* α for rejection (notice the *statistical reliability-sensitivity uncertainty relation)* or the ***p*-value**.
- The most general definition is:

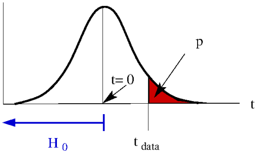
$$p = \text{Prob}(T \in E_{\text{data}} | H_0^*))$$

where the *extreme region* E_{data} contains all realisations of T that are further away from H_0 than t_{data}. Hence, t_{data} lies on the boundary of E_{data} Relation to the rejection region? p is defined such that $E_{\text{data}} = R(p)$

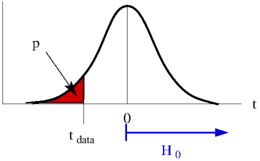
  - $p \geq 5\,\%$: not significant (no star at the value for β, sometimes a "+" if between 5 % and 10 %, e.g., $\beta_1 = 4.2^+$)
  - $p < 5\,\%$: significant (one star, e.g., $\beta_1 = 4.2^*$)
  - $p < 1\,\%$: very significant (two star, $\beta_1 = 4.2^{**}$)
  - $p < 0.001$: highly significant (three stars, $\beta_1 = 4.2^{***}$)

Calculating p for some basic tests

- Interval test $H_0 : \beta \leq \beta_0$ or $\beta < \beta_0$
  $p = P(T > t_{\text{data}} | \beta = \beta_0) = 1 - F_T(t_{\text{data}})$

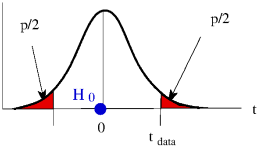


- Interval test $H_0 : \beta \geq \beta_0$ or $\beta > \beta_0$
  $p = P(T < t_{\text{data}} | \beta = \beta_0) = F_T(t_{\text{data}})$



- Point test $H_0 : \beta = \beta_0$ (symmetry of f_T assumed at the 3[rd] equality sign)

$$
\begin{aligned}
p &= P\big((T > |t_{\text{data}}|) \cup (T < -|t_{\text{data}}|)\big) \\
&= (1 - F_T(|t_{\text{data}}|)) + F_T(-|t_{\text{data}}|) \\
&= 1 - F_T(|t_{\text{data}}|) + 1 - F_T(|t_{\text{data}}|) \\
&= 2(1 - F_T(|t_{\text{data}}|))
\end{aligned}
$$



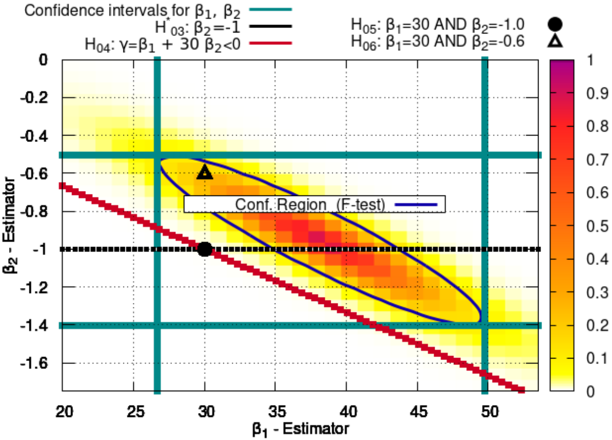
p-values for the null hypotheses of the hotel example

$$y = \beta_0 x_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

where $x_0 = 1$, x_1: proxy for quality [# stars]; x_2: price

- H_{01} "stars do not matter": point hypothesis $\beta_1 = 0$
  $t_{\text{data}} = 7.49,\ p = 2(1 - F_T^{(9)})(|t_{\text{data}}|) = 3.7E - 5^{***}$
- H_{02} "more stars are better": interval hypothesis $\beta_1 < 0$
  $t_{\text{data}} = 7.49,\ p = 1 - F_T^{(9)}(t_{\text{data}}) = 1.9E - 5^{***}$
- H_{03} "Δ occupancy $\le -1\,\%$ per addtl €": interval hypothesis $\beta_2 < -1$
  $t_{\text{data}} = 0.24,\ p = 1 - F_T^{(9)}(t_{\text{data}}) = 40\,\%$
- H_{04} One star more is worth less than 30 €":
  function interval hypothesis $\gamma = \beta_1 + 30\beta_2 < 0$
  $t_{\text{data}} = 4.20,\ p = 1 - F_T^{(9)}(t_{\text{data}}) = 0.12\,\%^{**}$
- H_{05} "star and price sensitivity simultaneously given":
  compound point hypothesis $(\beta_1 = 30) \cap (\beta_2 = -1)$
  $t_{\text{data}} = 11.8,\ p = 1 - F_F^{(2,9)}(t_{\text{data}}) = 0.30\,\%^{**}$

Visualization



- Turquoise lines: boundaries of the $\alpha = 5\,\%$-CIs of β_1 and β_2
- Black line: boundary of simple interval null hypothesis $H_{03} : \beta_2 \le -1$ (t-test)
- Red boxes: boundary of the function intervall hypothesis $H_{04} : \gamma = \beta_1 + 30\beta_2 < 0$ (t-test)
- Black symbols: simultaneous point hypotheses (F-test)
  •: $H_{05} : (\beta_1 = 30) \cap \beta_2 = -1)$, △: $H_{06} : (\beta_1 = 30) \cap (\beta_2 = -0.6)$.

4.2 Dependence on the True Parameter Value

All statistical tests, including the p-values, are based on some *null hypothesis* which is supposed to be *marginally* fulfilled, $\beta = \beta_0 \in H_0^*$. What if the true parameter values take on other values?

- Since regression parameters are continuous, the probability $P(H_0^*) = 0$ exactly, so the tests and p-values *do not reflect reality*
- What happens for other values $\beta \notin H_0^*$? This is quantified by following conditional probability called **statistical power function**:

$$\pi_\alpha(\beta) = \text{Pr(test rejected at error probability } \alpha|\beta)$$

- If $\beta \notin H_0$, then $\pi(\beta)$ indicates the **statistical power** or **specificity** of a test and $1 - \pi(\beta)$ its probability for a type-II error
- If $\beta \in H_0$, then $\pi(\beta)$ is the type-I (α) error and $1 - \pi(\beta)$ the **sensitivity** of a test
- Sensitivity and specificity depend on the assumed error probability α. By definition, $\pi(\beta_0) = \alpha$ if $\beta_0 \in H_0^*$

Calculating the statistical power function

- If $\beta \neq \beta_0 \in H_0^*$, then the usual test function, e.g., $(\hat{\beta}_j - \beta_{j0})/\sqrt{\hat{V}_{jj}}$ does *no longer* obey a standard statistical distribution such as standardnormal or student-t

- However, $T = (\hat{\beta}_j - \beta_j)/\sqrt{\hat{V}_{jj}}$ does:

$$T = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{V}_{jj}}} = \frac{\hat{\beta}_j - \beta_{j0}}{\sqrt{\hat{V}_{jj}}} + \frac{\beta_{j0} - \beta_j}{\sqrt{\hat{V}_{jj}}} = \frac{\hat{\beta}_j - \beta_{j0}}{\sqrt{\hat{V}_{jj}}} - \Delta T$$

- $\Rightarrow$ The independent variable of the power function is the standardized difference $\Delta T = (\beta_j - \beta_{j0})/\sqrt{\hat{V}_{jj}}$

Example I: Interval test for $<$ and $\leq$

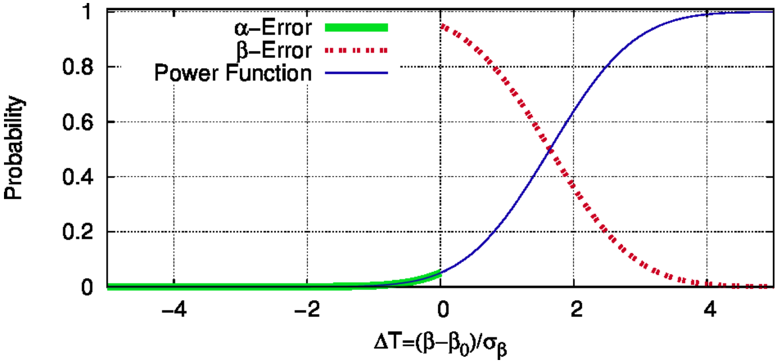
$$
\begin{aligned}
\pi^{\leq}(\Delta T) &\overset{\text{def rejection}}{=} P\left(\frac{\hat{\beta}_j - \beta_{j0}}{\sqrt{\hat{V}_{jj}}} > t_{1-\alpha}\right) \\
&\overset{\text{def } \Delta T}{=} P(T + \Delta T > t_{1-\alpha}) \\
&= P(T > -\Delta T + t_{1-\alpha}) \\
&= 1 - P(T < -\Delta T + t_{1-\alpha}) \\
&\overset{\text{symm.}}{=} P(T < \Delta T - t_{1-\alpha}) \\
&\overset{\text{def distr.}}{=} \underline{\underline{F_T(\Delta T - t_{1-\alpha})}}
\end{aligned}
$$

? Test this expression by calculating $\pi^{\leq}(0)$ and $\pi'^{\leq}(0)$

! Just insert $\Delta T = 0$:

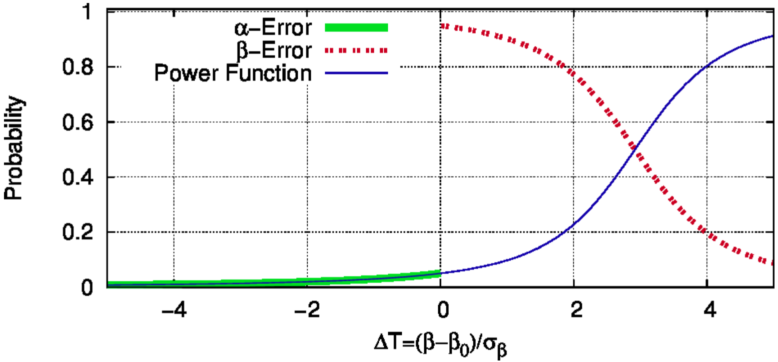
$$
\begin{aligned}
\pi^{\leq}(0) &= F_T(-t_{1-\alpha}) \\
&= F_T(t_\alpha) \\
&\overset{\text{def quantile}}{=} \alpha \quad \checkmark \\
\pi'^{\leq}(0) &= f_T(-t_{1-\alpha}) > 0 \quad \checkmark
\end{aligned}
$$

Type I and II errors for "<" or "≤"-tests as a function of the true value relative to H_0, known variance



- The maximum type-I error probability of α occurs if $\beta = \beta_0$, i.e., at the boundary of H_0.
- The maximum type-II error probability of $1 - \alpha$ occurs if β is just outside of H_0.

The same for unknown variance, df=2 degrees of freedom



- The increase with ΔT is steeper but $\pi(0) = \alpha$ is unchanged

Example II: Interval test for for $>$ and $\geq$

$$\pi^{\geq}(\Delta T) \overset{\text{def rejection}}{=} P\left(\frac{\hat{\beta}_j - \beta_{j0}}{\sqrt{\hat{V}_{jj}}} < t_\alpha\right)$$

$$\overset{\text{def } \Delta T}{=} P(T + \Delta T < t_\alpha)$$

$$= P(T < -\Delta T + t_\alpha)$$

$$\overset{\text{def distr.}}{=} \underline{\underline{F_T(t_\alpha - \Delta T)}}$$

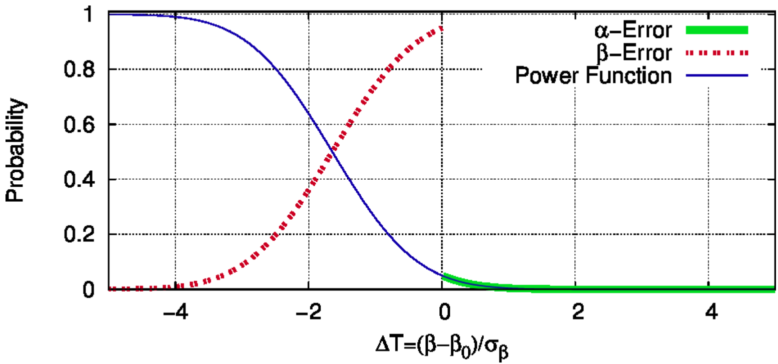
? Test this expression by calculating $\pi^{\geq}(0)$ and $\pi'^{\geq}(0)$

! Just insert $\Delta T = 0$:

$$\pi^{\geq}(0) \overset{\text{def quantile}}{=} \alpha \; ✓$$

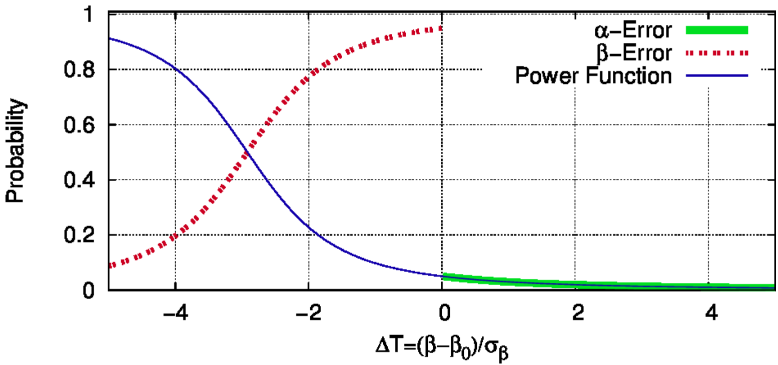
$$\pi'^{\geq}(0) = -f_T(0) < 0 \; ✓$$

Type I and II errors for ">" or "≥"-tests, known variance



- Again, the maximum type I and II error probabilities of α and $1-\alpha$, respectively, are obtained if the true parameter(s) are at the boundary / very near outside of H_0.
- The maximum type-I error probability is also known as significance level.

The same for unknown variance, df=2 degrees of freedom



Example III: Point test for "="

$$\pi^{\text{eq}}(\Delta T) \overset{\text{def rejection}}{=} P\left(\left|\frac{\hat{\beta}_j - \beta_{j0}}{\hat{\sigma}_{\hat{\beta}_j}}\right| > t_{1-\alpha/2}\right)$$

$$\overset{\text{def } \Delta T}{=} P(|T + \Delta T| > t_{1-\alpha/2})$$

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

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

$$\overset{\text{def distr.}}{=} 1 - F_T(t_{1-\alpha/2} - \Delta T) + F_T(-t_{1-\alpha/2} - \Delta T)$$

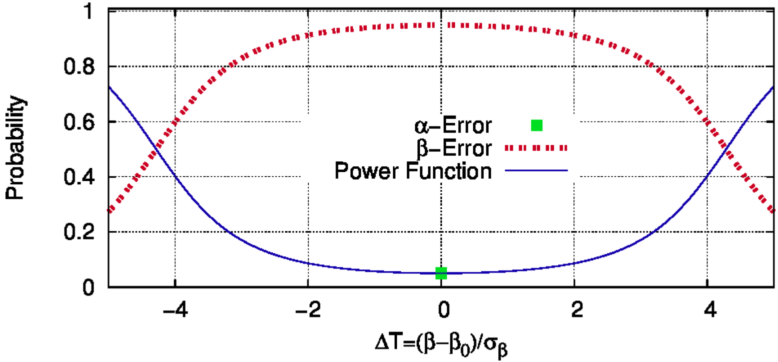
$$\overset{\text{symm.}}{=} \underline{\underline{2 - F_T(t_{1-\alpha/2} - \Delta T) - F_T(t_{1-\alpha/2} + \Delta T)}}$$

? Test this expression by calculating $\pi^{\le}(0)$

! Just insert $\Delta T = 0$:

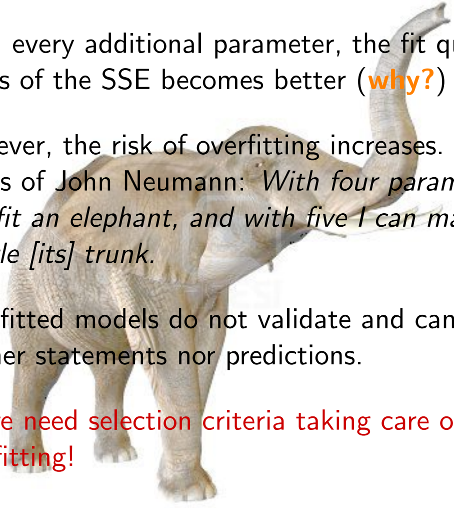
$$\pi^{\text{eq}}(0) = 2 - (1 - \alpha/2) - (1 - \alpha/2) = \alpha \checkmark$$

Type I and II errors for two-sided (point-)tests (unkown variance, df=2)



- Since H_0 is a point set here, the type-I error probability is always given by α ("significance level")

4.3 Model Selection Strategies
Problem Statement



- With every additional parameter, the fit quality in terms of the SSE becomes better (**why?**)
- However, the risk of overfitting increases. In the words of John Neumann: *With four parameters I can fit an elephant, and with five I can make him wiggle [its] trunk.*
- Overfitted models do not validate and can make neither statements nor predictions.
- $\Rightarrow$ we need selection criteria taking care of overfitting!

Model selection: some standard criteria

- **(1) Adjusted R^2:**

$$\bar{R}^2 = 1 - \frac{n-1}{n-J-1}\left(1-R^2\right), \quad R^2 = 1 - \frac{S}{S_0},$$

$S =$ SSE(calibr. full model), $S_0 =$ SSE(calibr. constant-only model).

- **(2) Akaike information criterion AIC:**

$$\text{AIC} = \ln \hat{\sigma}^2_{\text{descr}} + J\,\frac{2}{n},$$

- **(3) Bayes' Information criterion BIC:**

$$\text{BIC} = \ln \hat{\sigma}^2_{\text{descr}} + J\,\frac{\ln n}{n}\,.$$

Notice that the descriptive $\hat{\sigma}^2_{\text{descr}} = S/n$ instead of the unbiased $\hat{\sigma}^2 = S/(n-1-J)$ are assumed when defining AIC and BIC.

Model selection: Strategy à la "Occam's Razor"

- Identify J possibly relevant exogenous factors (the constant is always included) and calculate $\bar{R}^2$, AIC, or BIC for all 2^J combinations of these factors (a given factor is either contained or not) by *brute force*).
- The best model is that maximizing $\bar{R}^2$ or minimizing AIC or BIC.
- Since AIC and also $\bar{R}^2$ penalize complex models (with many parameters) too little, the BIC is usually the best bet.
- Besides the *brute-force* approach, there are two faster strategies that may not find the "best" model (BIC etc are not transitive)
  - **Top-down approach**: Start with all the J factors. In each round, eliminate a single factor such that the reduced model has the highest increase in $\bar{R}^2$ / decrease in AIC or BIC. Stop if there is no further improvement.
  - **Bottom-up approach**: Start with the constant-only model $y = \beta_0$ and successively add factors until there is no further improvement.
- Standard statistics packages contain all of these strategies.

4.4. Logistic regression

- Normal linear models of the form $Y = \boldsymbol{\beta}'\boldsymbol{x} + \epsilon$ require the endogenous variable to be continuous (discuss!)
- Using model chaining with an unobservable intermediate continuous variable Y^* allows one to model binary outcomes:

$$Y(\boldsymbol{x}) = \begin{cases} 1 & Y^*(\boldsymbol{x}) > 0 \\ 0 & \text{otherwise,} \end{cases} \quad Y^*(\boldsymbol{x}) = \hat{y}^*(\boldsymbol{x}) + \epsilon = \boldsymbol{\beta}'\boldsymbol{x} + \epsilon$$

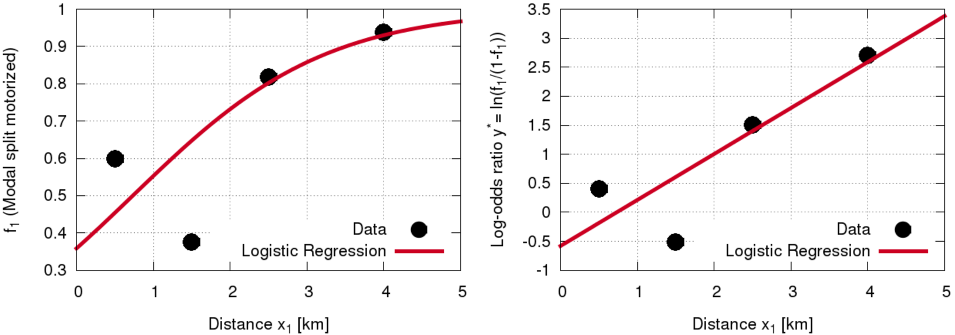
  where ϵ obeys the **logistic distribution** with $F_\epsilon(x) = e^x/(e^x + 1)$
- Probability P_1 for the outcome $Y = 1$:

$$P_1 = P(Y^*(\boldsymbol{x}) > 0) = F_\epsilon(\boldsymbol{\beta}'\boldsymbol{x}) = \frac{e^{\boldsymbol{\beta}'\boldsymbol{x}}}{e^{\boldsymbol{\beta}'\boldsymbol{x}} + 1}$$

- Formally, this is a normal linear regression model for the log of the **odds ratio** $P_1/P_0 = P1/(1 - P_1)$:

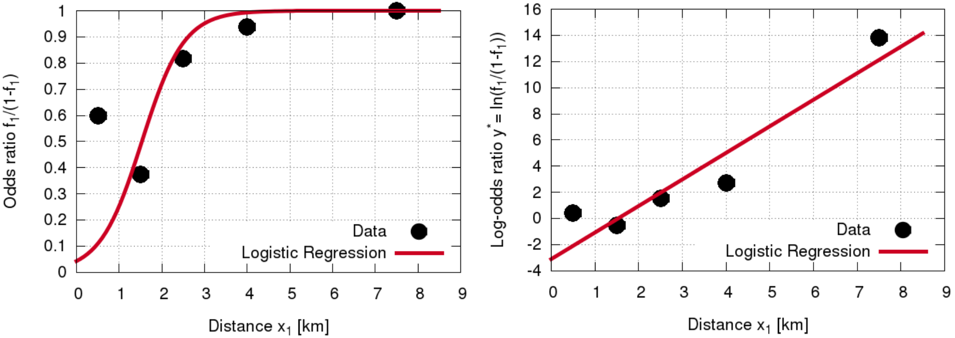
$$\hat{y}^*(\boldsymbol{x}) = \boldsymbol{\beta}'\boldsymbol{x} = \ln\left(\frac{P_1}{P_0}\right)$$

Example: naive OLS-estimation (RP student interviews)



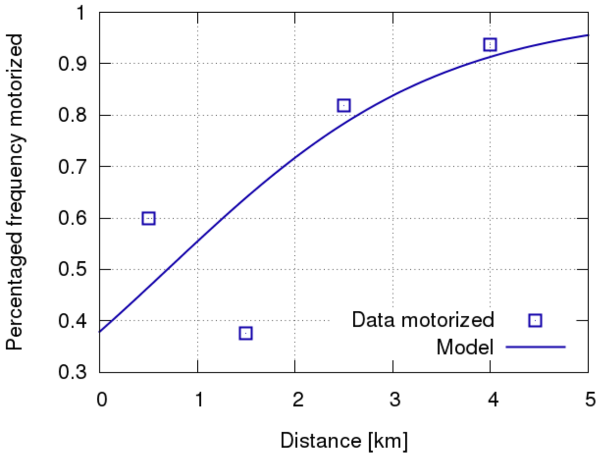
- Alternatives: $i = 1$: motorized and $i = 2$ (not)
- Intermediate variable estimated by percentaged choices: $y^* = \ln(f_1/(1 - f_1))$
- Model: Log. regression, $\hat{y}^*(x_1) = \beta_0 + \beta_1 x_1$
- OLS Estimation: $\beta_0 = -0.58, \quad \beta_1 = 0.79$

Method consistent? added 5th data point with f=0.9999



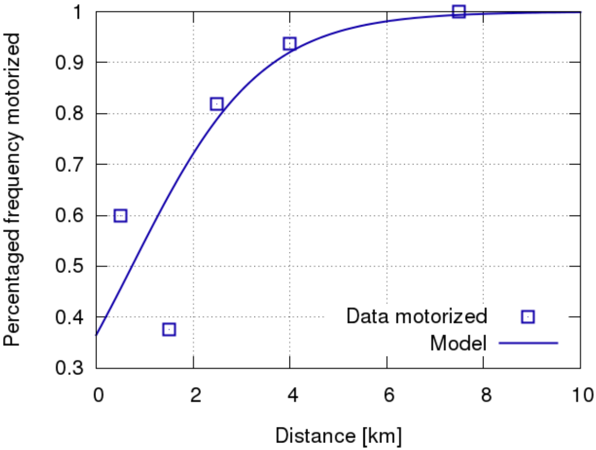
- Same model: $\hat{y}^*(x_1) = \beta_0 + \beta_1 x_1$
- New estimation: $\beta_0 = -3.12, \quad \beta_1 = 2.03$
- Estimation would fail if $f_1 = 0$ or $=1$ $\Rightarrow$ real discrete-choice model necessary!

Comparison: real Maximum-Likelihood (ML) estimation



- Model: Logit, $V_i(x_1) = \beta_0\delta_{i1} + \beta_1 x_1 \delta_{i1}$, $V_2 = 0$.
- Estimation: $\beta_0 = -0.50 \pm 0.65$, $\beta_1 = +0.71 \pm 0.30$

Comparison: real ML estimation with added 5th data point



- Same logit model, $V_i(x_1) = \beta_0 \delta_{i1} + \beta_1 x_1 \delta_{i1}$, $V_2 = 0$.
- New estimation: $\beta_0 = -0.55 \pm 0.63$, $\beta_1 = +0.75 \pm 0.27$