Inferential Statistics: Estimation and Testing

Tsvetana Spasova

Summer School on Modeling, AI, and Complex Systems

Gyulechica, June 2024

Contents

Objectives

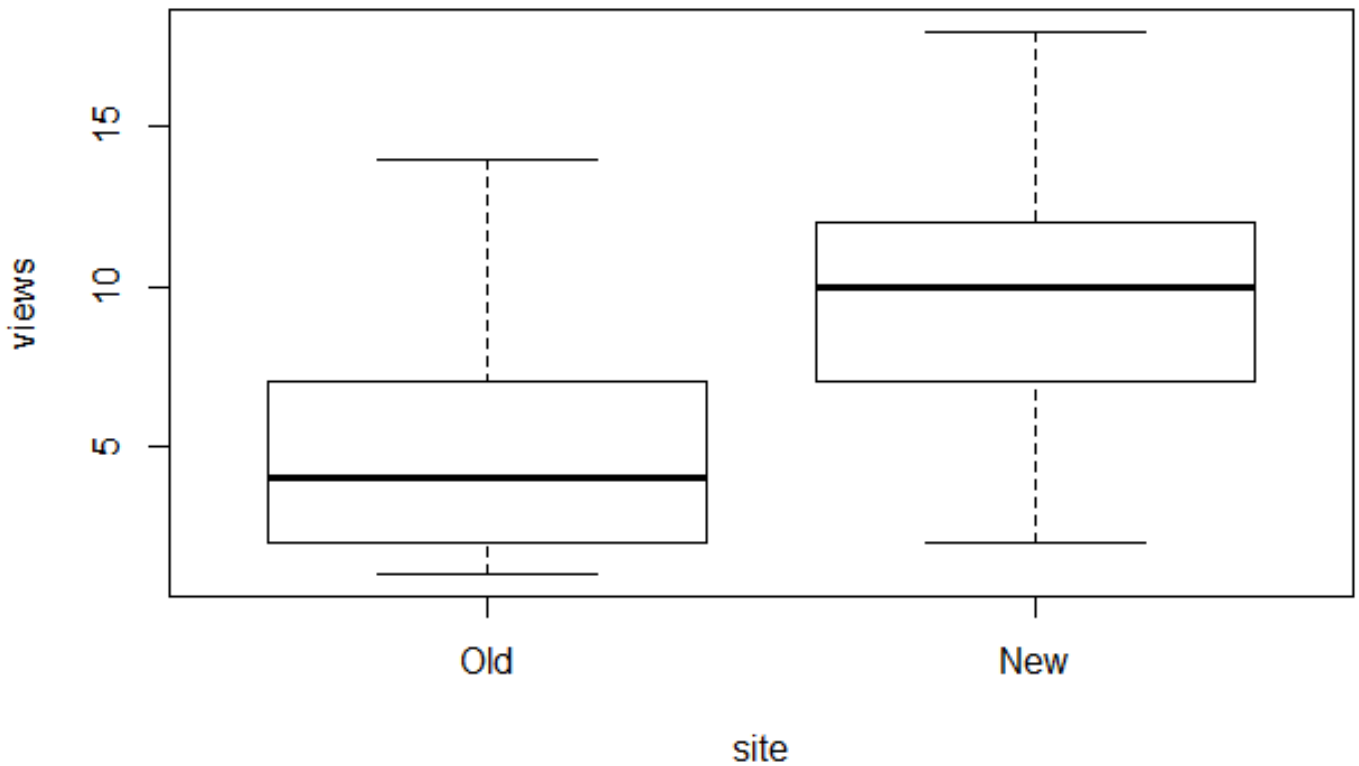
- Understand estimation.
- Be able to calculate and interpret a confidence interval for a proportion and for a mean.
- Understand hypothesis testing.
- Be able to calculate and interpret one sample and two-sample tests.

SDV2015[1] Chapters 9, 11.1-11.4, 12, 13 and 14

[1] Sharpe, N.R., De Veaux, R.D., Velleman, P.F. (2015) Business Statistics (3ed.), Global Edition, Pearson.

Old and new website: pageviews

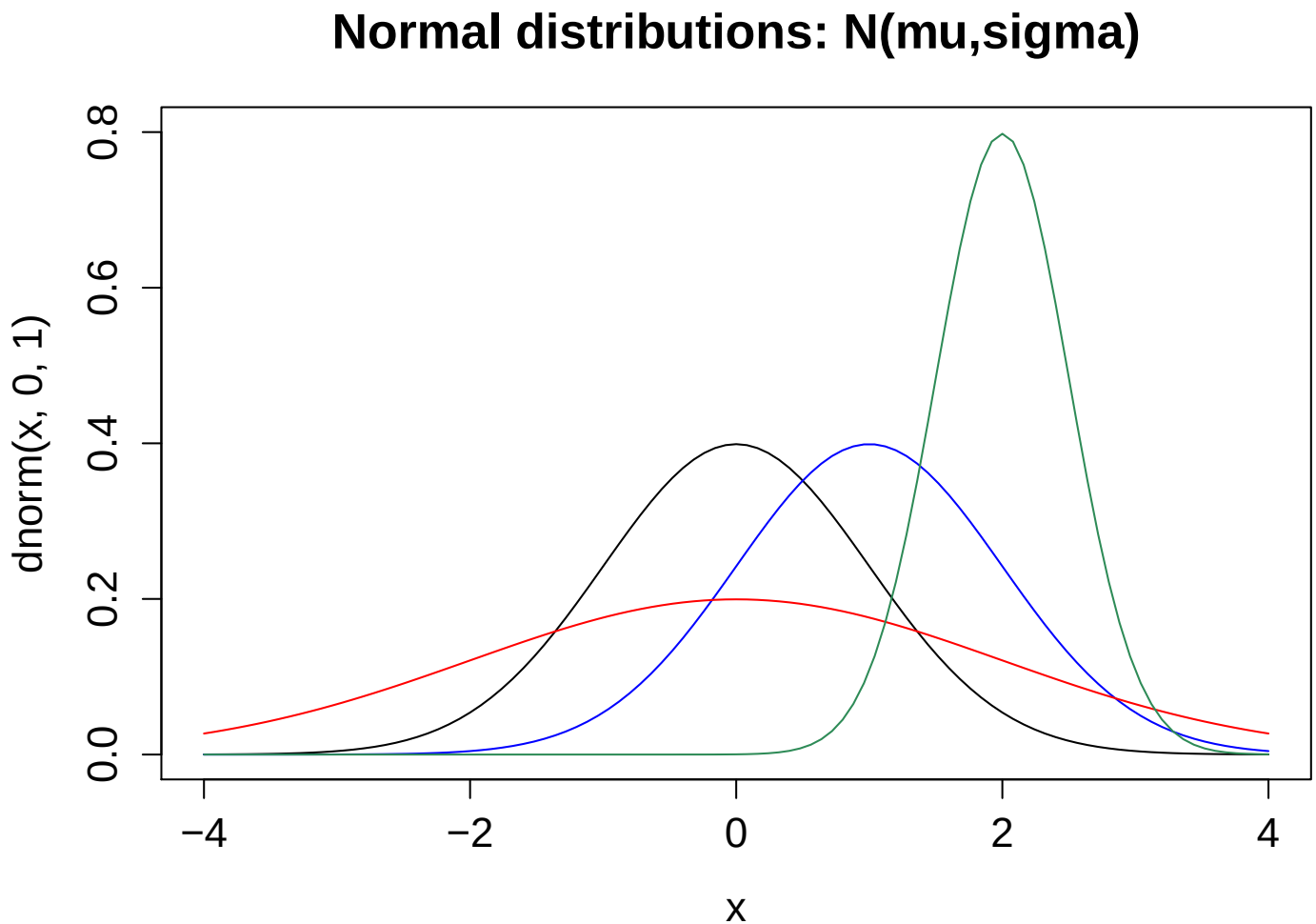
Old website: $9, 4, 14, 7, 2, 3, 5, 1, 2, 4$
New website: $10, 13, 2, 3, 7, 9, 11, 18, 10, 12$



Random variables

- The uncertainty in a measurement is conceptualized as a stochastic mechanism called **random variable** (r.v.).
- A random variable is a random experiment that has a real number, or one of a defined number of categories, as its outcome. The probabilities of the outcomes form the **distribution** of the r.v..
- For continuous r.v. the distribution is characterised with a **probability density function** (pdf).
- For **parametric** distributions the pdf is known up to some unknown parameters, e.g. the normal distribution with unknown location μ and unknown standard deviation σ: $f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$.

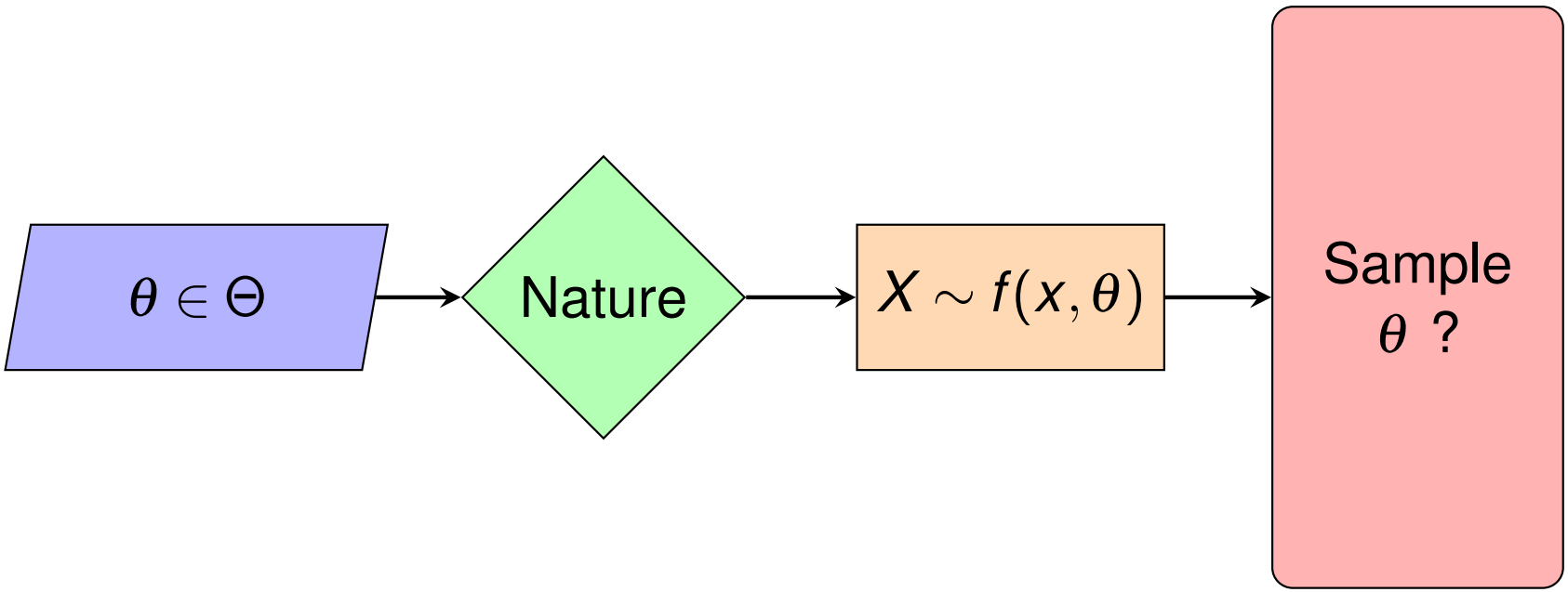
Normal distribution $N(\mu, \sigma)$



Sampling

- Sample: Set of observations indexed by $i = 1, \ldots, n$.
- Measurements on observation i: x_i, y_i, z_i etc.
- "Similar" measurements are to be expected: Assume that a r.v. X has been realised n times **independently** and resulted in the sample values $x_i, i = 1, \ldots, n$.
- Assume X has a known probability distribution except for a parameter θ: $X \sim f(x; \theta)$ (The tilde "$\sim$" means "distributed according to").
- E.g. assume that the pageviews on the old website stems from independent, identically distributed realisations of a normal r.v. $N(\mu, \sigma)$.

Estimation



Estimation ctd.

- The general form of the distribution, say pdf $f(x;\theta)$ is assumed known. Note that there are infinitely many possible distribution forms!
- The statistician must check whether this model of the reality is good enough! E.g. using a normal plot.
- Independence is crucial because it ensures, that with every new observation of the assumed random variable we get new additional information. Otherwise variance estimators are biased and tests are misleading!

Mean estimation

- The sample mean of the observations $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ estimates μ and the sample standard deviation $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ estimates σ.
- Note that for another sample the sample mean $\bar{x}$ and the sample standard deviation*s* would be different!
- If $X_i \sim N(\mu, \sigma)$ then

$$\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

- The standard error of the mean, $SEM = s/\sqrt{n}$, estimates $\sigma/\sqrt{n}$.

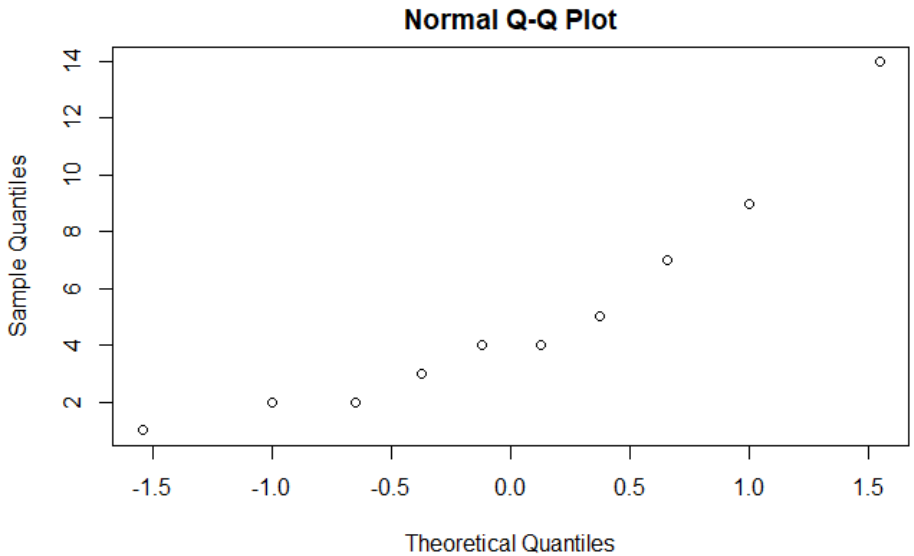
Confidence interval

- A confidence interval is calculated from the sample and covers the intended parameter with specified probability.
- If $X_i \sim N(\mu, \sigma)$ then

$$\bar{X} \pm t \cdot SEM$$

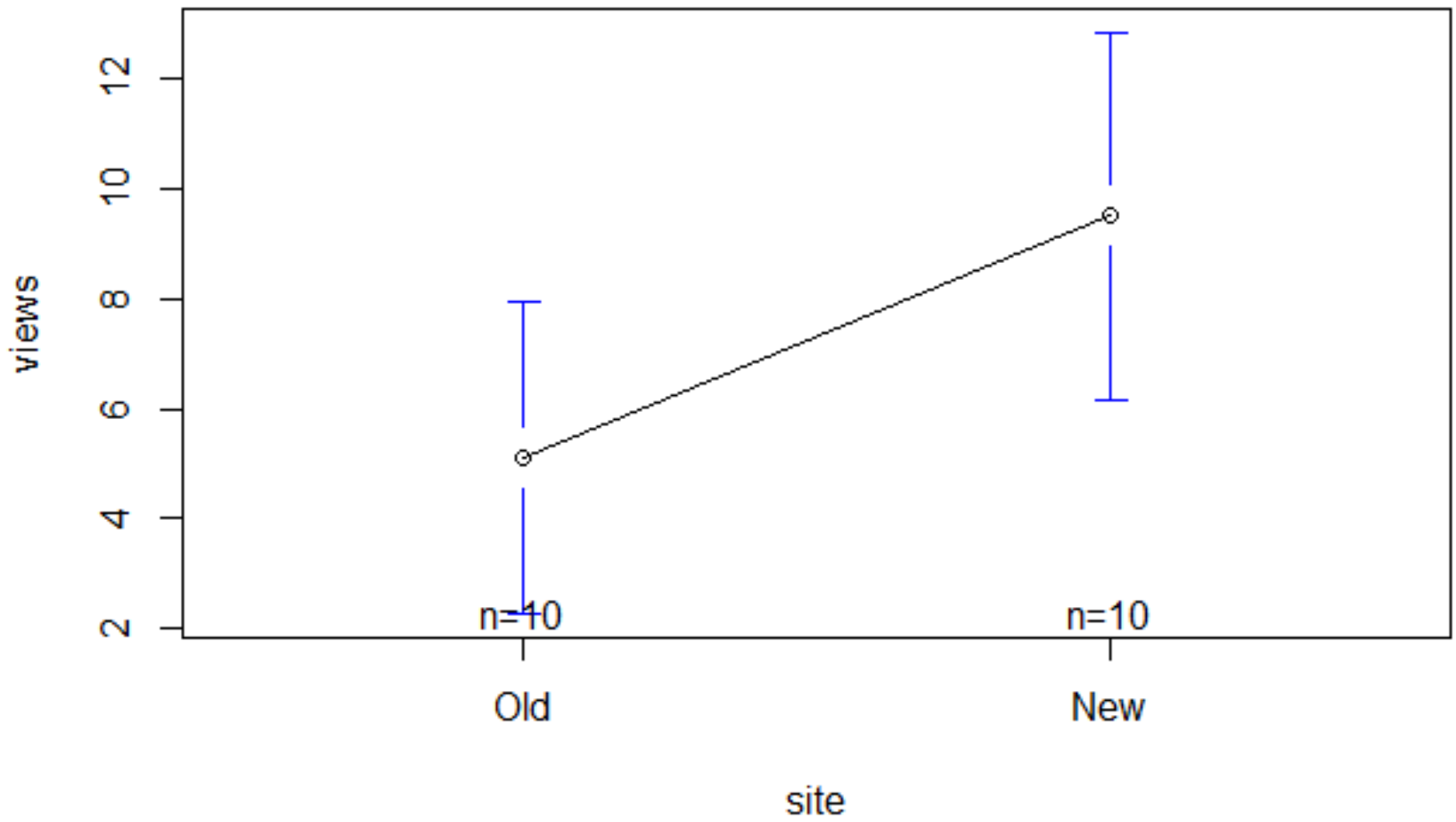
  is a confidence interval. Here t is a quantile of the t-distribution with $n-1$ degrees of freedom.
- For a 95%-confidence interval $t = t(0.975, n-1)$. E.g. for $n = 26$ the factor t is $t(0.975, 25) = 2.06$.
- If X_i has only approximately a normal distribution then the above confidence is still approximately correct thanks to the **Central Limit Theorem**.

95% CI for pageviews on old website



- $\bar{x} = 5.1$, $s = 3.957$, $SEM = s/\sqrt{n} = 3.957/\sqrt{10} = 1.251$
- $t = t(0.975, 9) = 2.262$
- 95% CI: $5.1 \pm 2.262 \cdot 1.251 = [2.269, 7.931]$.

Pageviews with confidence interval



95% CI for a proportion

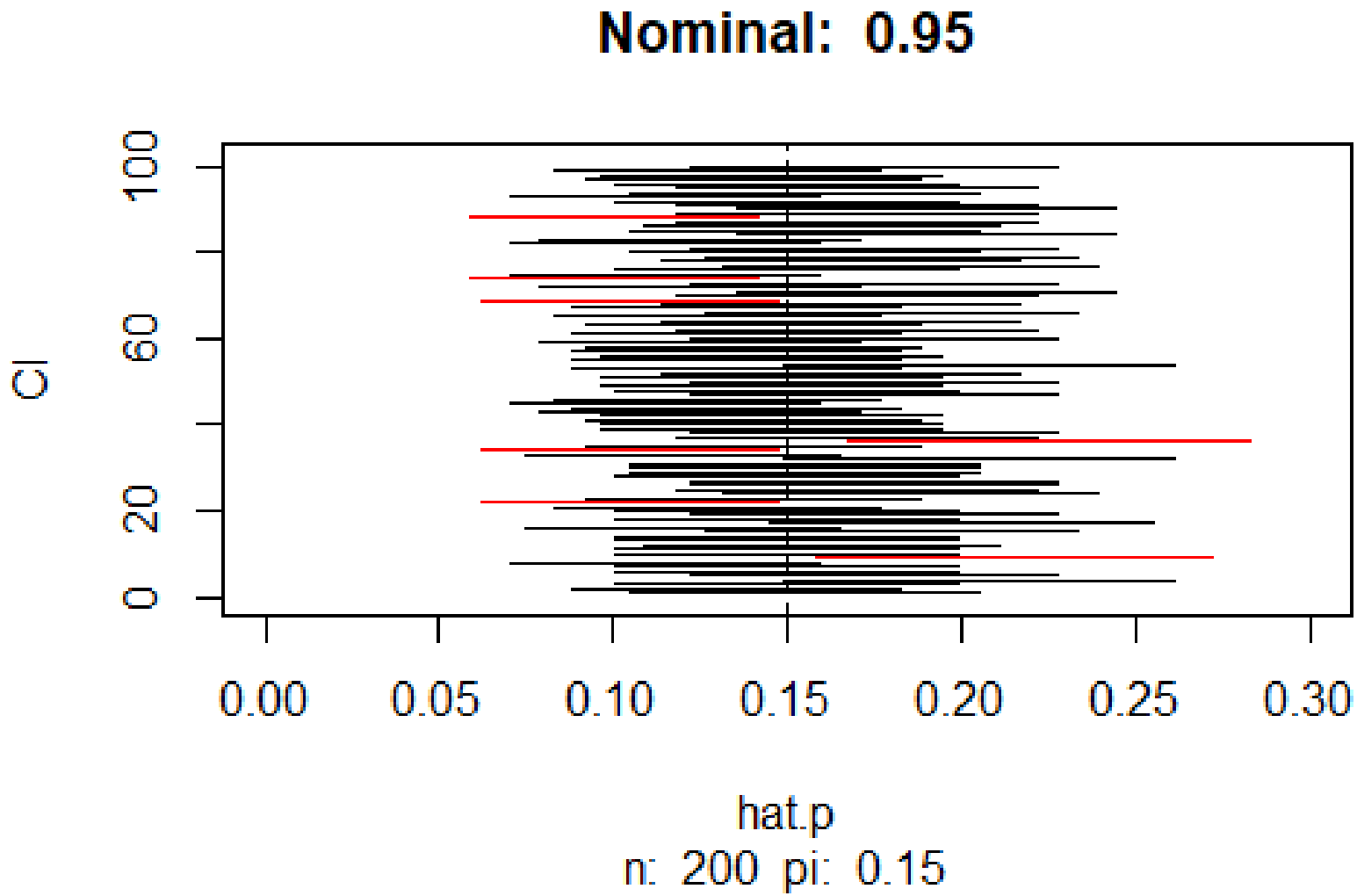
- Suppose the 200 visits on the old website yielded 24 sales (conversions).
- The sample proportion of conversions is $p = 24/200 = 0.12$.
- The standard error of proportion depends directly on p: $SEP = \sqrt{p \cdot (1-p)/n}$.
- A confidence interval for the true proportion of conversions in many more visits is

$$p \pm z \cdot SEP.$$

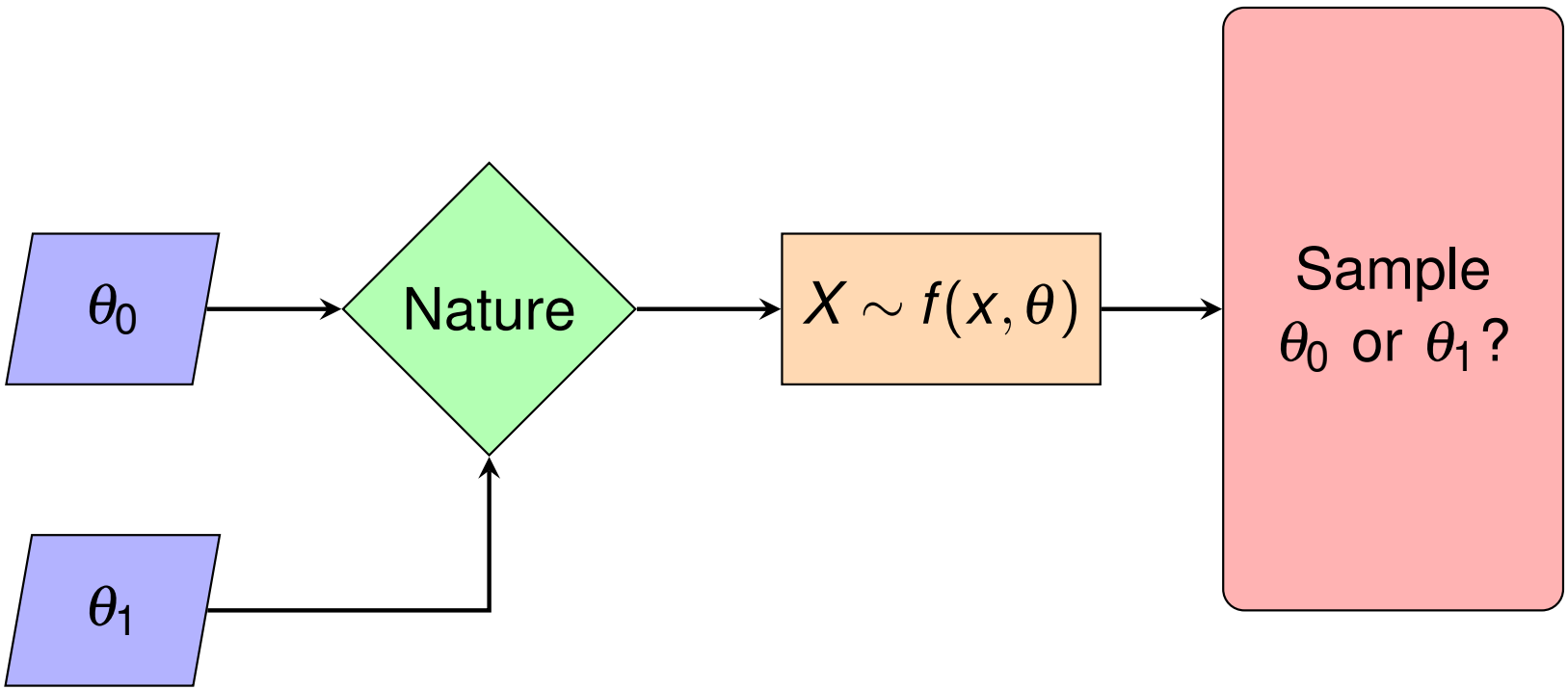
  Here z is a quantile of the standard normal distribution.
- For a 95% confidence interval $z = z(0.975) = 1.96$.
- A 95% confidence interval for the true proportion of conversions is

$$p \pm z \cdot SEP = 0.12 \pm 1.96 \cdot 0.023 = [0.075, 0.165]$$

Simulation of confidence intervals



Testing



Testing for proportion of conversions

Question: Is the conversion rate higher than $\pi_0 = 0.1$?

1. Assumption: web-visits are independent realisations of a Bernoulli r.v. with equal conversion rate π.
2. $H_0 : \pi \leq \pi_0, H_1 : \pi > \pi_0$ with $\pi_0 = 0.1$
3. Significance level α, e.g. $\alpha = 0.05$.
4. Test statistic and distribution:

$$Z = \frac{p - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}} \sim N(0,1)$$

   (standard normal distribution) if $n\pi_0 > 10$ and $n(1-\pi_0) > 10$.
5. Observed test statistic for $p = 0.12$ is $z_{obs} = 0.943$.
6. Probability to be more extreme than z_{obs} (p-value): $pv = P[Z > z_{obs}] = 0.173$.
7. If $pv < \alpha$ reject H_0. Here $0.173 \not< 0.05$: cannot reject H_0!

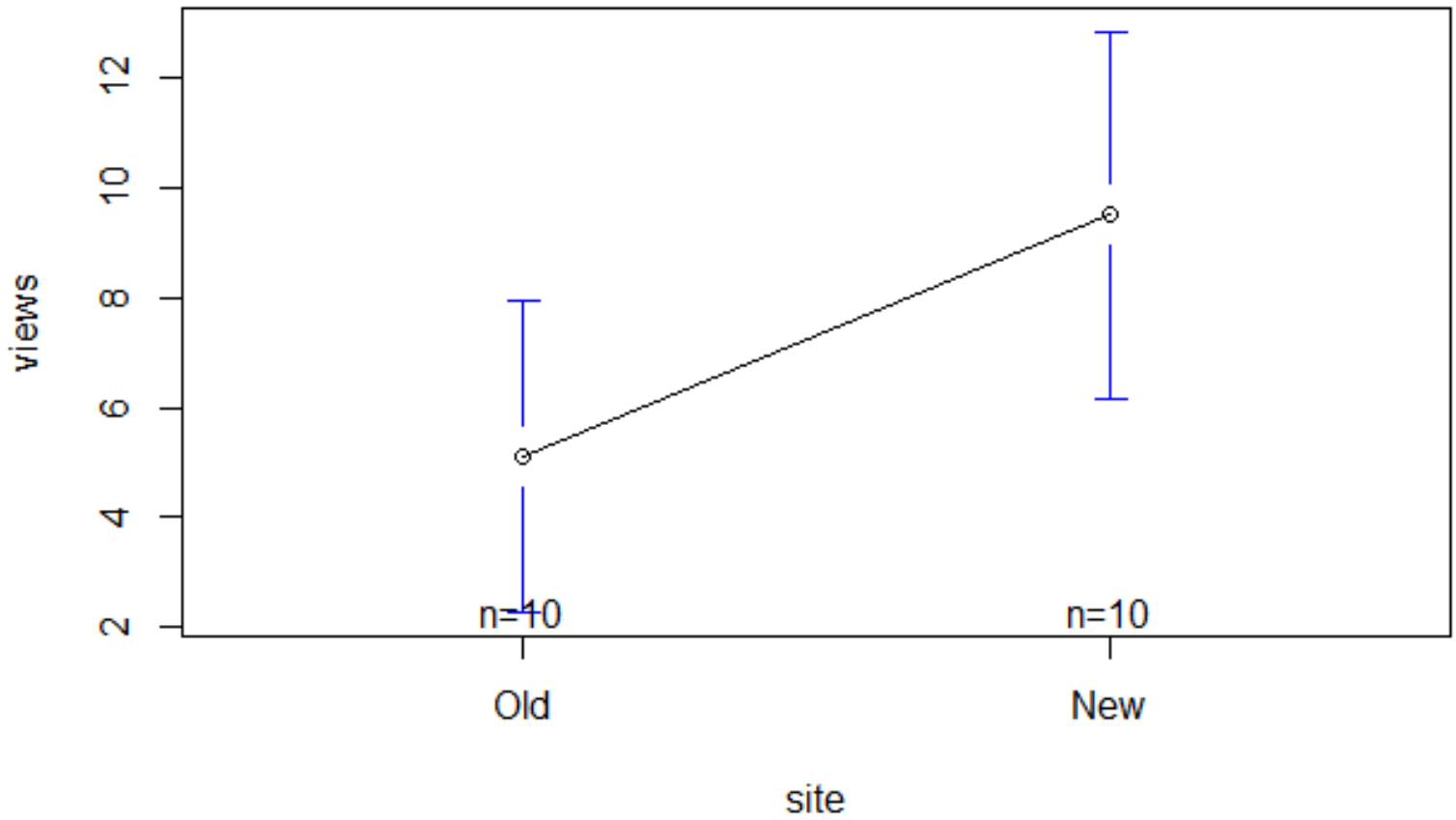
One-sample t-test (one-sided)

1. $X_1, \ldots, X_n$ i.i.d. $\sim N(\mu, \sigma^2)$
2. $H_0 : \mu \leq \mu_0$, $H_1 : \mu > \mu_0$ (one-sided alternative)
3. Fix significance level α
4. Fix test statistic

$$T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}.$$

5. Calculate observed test statistic t_{obs}.
6. Calculate p-value $pv = P_{H_0}[T > t_{obs}]$
7. If $pv \leq \alpha$ reject H_0 otherwise retain H_0.

Pageviews with confidence interval



Are the means of the two websites actually equal?

Pooled two-sample t-test

- Two independent samples $x_{11}, \ldots, x_{1n_1}$ and $x_{21}, \ldots, x_{2n_2}$.
- Approximately $X_{1i} \sim N(\mu_1, \sigma_1)$ and $X_{2i} \sim N(\mu_2, \sigma_2)$.
- Equal variances: $\sigma_1 = \sigma_2$.
- $H_0 : \mu_1 - \mu_2 = \Delta_0$ (usually $\Delta_0 = 0$). $H_1 : \mu_1 - \mu_2 \neq \Delta_0$.
- Test statistic

$$T = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

  where

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

  is the pooled variance.
- Distribution $T \sim t_{n_1+n_2-2}$; p-value = $P_{H_0}[|T| > |t_{obs}|]$
- p-value small? fix level of significance α, often $\alpha = 0.05$ or $\alpha = 0.01$ (before the test!). Decision: if **p-value** $< \alpha$ **reject** H_0

Wilcoxon-Mann-Whitney U-test

- Two independent samples $x_{11}, \ldots, x_{1n_1}$ and $x_{21}, \ldots, x_{2n_2}$, $X_{1i} \sim F_2$ and $X_{2i} \sim F_2$
- H_0: $F_1 = F_2$, $H_1 : F_2(x) = F_1(x + \Delta)$ (shifted).
- Rank the two samples jointly and let W_1 be the sum of the ranks of the $x_{1i}, i = 1, \ldots, n$.
- Test statistic

$$T = \frac{W_1 - n_1 \bar{n}}{\sqrt{n_1 n_2 \bar{n}/6}},$$

  where $\bar{n} = (n_1 + n_2 + 1)/2$.
- T has an approximate standard normal distribution if $n_1 > 8$ and $n_2 > 8$.

Contingency table: Rooms vs New Real Estate (NRE)

Rooms	NRE=0	NRE=1	Total
1	18	5	23
2	23	6	29
3	32	5	37
4	14	18	32
5	14	17	31
Total	101	51	152

Is the number of rooms independent of NRE?

Denote the count in row i and column j with n_{ij} and row-sums as $n_{i.}$, column-sums $n_{.j}$ total sum n.

Prediction under the independence model

Rooms	NRE=0	NRE=1	Total
1			23
2		9.73	29
3			37
4			32
5			31
Total	101	51	152

- Independence: $\pi_{ij} = \pi_i \cdot \pi_j$ (multiplicative law).
- Expected count

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

- Expected count for 2 room appartments of NRE: $\hat{n}_{22} = 152 \cdot \frac{29}{152} \frac{51}{152} = 9.73$.

χ^2-test for contingency tables

- R rows and C columns.
- Null hypothesis $H_0: \ \pi_{ij} = \pi_i \cdot \pi_j$.
- Test statistic (chi-squared statistic, χ^2-statistic)

$$T = \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}$$

- Approximate distribution $T \sim \chi^2_{(R-1)(C-1)}$ if no $\hat{n}_{ij} < 5$.
- $p - value = P_{H_0}[T > t_{obs}] < \alpha$ reject H_0.

χ^2 ctd.

- If a chi-squared test is significant it is interesting to see, which cells contribute most to the χ^2-statistic:
  - **Residuals**: $r_{ij} = (n_{ij} - \hat{n}_{ij})/\sqrt{\hat{n}_{ij}}$
  - Approximately $r_{ij} \sim N(0,1)$.
- And it is interesting to see how large the **association** between the two involved variables is:
  - Cramer's V:

$$V = \sqrt{\frac{\chi^2_{obs}}{n \cdot \min(R-1, C-1)}}$$

  - $0 < V < 1$
  - strength of association
    $V < 0.1$: weak association
    $0.1 \leq V < 0.3$: moderate association
    $V \geq 0.3$: strong association

Keep in mind!

- A model is not the reality!
- Every method has assumptions: Your data will not meet these assumptions fully!
- A statistical method **quantifies** uncertainty taking into account the available information: Every estimator has a standard error!
- A statistical test helps to decide rationally under uncertainty.
- Significance $\neq$ Relevance

Software for Statistics (=Analytics)

- Don't use Excel for data analysis!
- **R** is the most flexible and advanced statistical software with a large number of packages for statistical methods and fantastic graphing capabilities (`https://www.r-project.org/`)
- **Python** is a programming language with a formal definition (`https://www.python.org/`)
- **SAS** and **SPSS** are well established statistical software in the business world. Both have interfaces to R.