Université de Montréal

Fiche Récapitulative

Statistique

Julien Hébert-Doutreloux

April 26, 2020

Contents

1 Chapitre 4

1.1 Règles pour les moyennes

Théorème 1. *Si X est une variable aléatoire et a et b des constantes, alors*

$$E(a+bX) = \mu_{a+bX} = a + b\mu_X = a + bE(X) \tag{1}$$

Si X et Y sont des variables aléatoires (indépendantes ou non), alors

$$E(X+Y) = \mu_{X\pm Y} = \mu_X \pm \mu_Y = E(X) \pm E(Y) \tag{2}$$

1.2 Règles pour les variances

Définition 1. *(Variance) Soit $x_1, x_2, ...$, les événements élémentaires de la variable aléatoire discrète X et $p_1, p_2, ...$, leur probabilité respective. La variance de X est donnée par*

$$\sigma_X^2 = \sum_{i=1}^{\infty}(x_i - \mu_X)^2 p_i = E(X-\mu_X)^2$$

L'écart type est la racine carrée de la variance.

Théorème 2. *Si X est une variable aléatoire et a et b des constantes, alors*

$$\sigma_{a+bX}^2 = b^2\sigma_X^2 \tag{3}$$

Si X et Y sont des variables aléatoires indépendantes, alors

$$\sigma_{Y\pm X}^2 = \sigma_X^2 + \sigma_Y^2 \tag{4}$$

Si X et Y ont une corrélation de ρ, alors

$$\sigma_{Y\pm X}^2 = \sigma_X^2 + \sigma_Y^2 \pm 2\rho\sigma_X\sigma_Y \tag{5}$$

1.3 Règles générales des probabilités

Proposition 1.

Règle d'addition pour l'union de deux événements

$$P(A\cup B) = P(A \text{ ou } B) = P(A) + P(B) - P(A\cap B) \tag{6}$$

Probabilité conditionnelle

$$P(B|A) = \frac{P(A\cap B)}{P(A)} \tag{7}$$

Règle de multiplication pour l'intersection

$$P(A\cap B) = P(A)P(B|A) = P(B)P(A|B) \tag{8}$$
$$P(A\cap B\cap C) = P(A)P(B|A)P(C|A\cap B) \tag{9}$$

Théorème de Bayes

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^C)P(A^C)} \tag{10}$$

Indépendance (soit A et B indépendant, alors)

$$P(B|A) = P(B) \tag{11}$$
$$P(A\cap B) = P(A)P(B) \tag{12}$$

2 Chapitre 5: Sampling Distributions

2.1 5.1 Toward Statistical Inference

- A number that describes a population is a **parameter**. A number that describes a sample (is computed from the sample data) is a **statistic**. The purpose of sampling or experimentation is usually **inference**: use sample statistics to make statements about unknown population parameters.
- A statistic from a probability sample or a randomized experiment has a **sampling distribution** that describes how the statistic varies in repeated data productions. The sampling distribution answers the question "What would happen if we repeated the sample or experiment many ?" Formal statistical inference is based on the sampling distributions of statistics.
- A statistic as an estimator of a parameter may suffer from **bias** or from high **variability**. Bias means that the center of the sampling distribution is not equal to the true value of the parameter. The variability of the statistic is described by the spread of its sampling distribution. Variability is usually reported by giving a **margin of error** for conclusions based on sample results.
- Properly chosen statistics from randomized data production designs have no bias resulting from the way the sample is selected or the way the experimental units are assigned to treatments. We can reduce the variability of the statistic by increasing the size of the sample or the size of the experimental groups.

2.2 5.2 The samplig Distributions of a Sample Mean

- The **population distribution** of a variable is the distribution of its values for all members of the population.
- The **sample mean** $\bar{X}$ of an SRS of size n drawn from a large population with mean μ and standard deviation σ has a sampling distribution with mean and standard deviation

$$\mu_{\bar{X}} = \mu$$
$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

  The sample mean $\bar{X}$ is an **unbiased estimator** of the population mean μ and is less variable than a single observation. The standard deviation decreases in proportion to the square root of the sample size n. This means that to **reduce** the standard deviation by a factor of C, we need to **increase** the sample size by a factor of C^2.
- The **central limit theorem** states that, for large n, the sampling distribution of $\bar{X}$ is approximately $N(\mu, \sigma/\sqrt{n})$ for any population with mean μ and finite standard deviation σ. This allows us to approximate probability calculations of $\bar{X}$ using the Normal distribution.
- Linear combinations of independent Normal random variables have Normal distributions. In particular, if the population has a Normal distribution, so does $\bar{X}$.

2.3 5.3 Sampling Distributions for Counts and Proportions

- A **count** X of successes has the **binomial distribution** $B(n, p)$ in the **binomial setting**: there are n trials, all independent, each resulting in a success or a failure, and each having the same probability p of a success.
- The binomial distribution $B(n, p)$ is a good approximation to the **sampling distribution of the count of successes** in an *SRS* of size n from a large population containing proportion p of successes. *We will use this approximation when the population is at least* 20 *- larger than the sample.*
- The **sample proportion** of successes $\hat{p} = X/n$ is an estimator of the population proportion p. It does not have a binomial distribution, but we can do probability calculations about $\hat{p}$ by restating them in terms of X.

- **Binomial probabilities** are most easily found by software. There is an exact formula that is practical for calculations when n is small. **Table C** contains binomial probabilities for some values of n and p. *For large n, you can use the Normal approximation.*

- The mean and standard deviation of a **binomial count** X and a **sample proportion** $\hat{p} = X/n$ are

$$\begin{array}{cc} \mu_X = np & \mu_{\hat{p}} = p \\ \sigma_X = \sqrt{np(1-p)} & \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} \end{array}$$

  The sample proportion $\hat{p}$ is, therefore, an **unbiased estimator** of the population proportion p.

- The **Normal approximation** to the binomial distribution says that if X is a count having the $B(n, p)$ distribution, then when n is large,

$$\begin{array}{c} X \text{ is approximately } N\left(np, \sqrt{np(1-p)}\right) \\ \hat{p} \text{ is approximately } N\left(p, \sqrt{\frac{p(1-p)}{n}}\right) \end{array}$$

  We will use this approximation when $np \geq 10$ and $n(1-p) \leq 10$. It allows us to approximate probability calculations about X and $\hat{p}$ using the Normal distribution.

- The **continuity correction** improves the accuracy of the Normal approximations.

- The exact **binomial probability formula** is

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

  where the possible values of X are $k = 0, 1, ..., n$. The binomial probability formula uses the **binomial coefficient**

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

- Here the factorial $n!$ is

$$n! = n \times (n-1) \times (n-2) \times ... \times 3 \times 2 \times 1$$

  for positive whole numbers n and $0! = 1$. The binomial coefficient counts the number of ways of distributing k successes among n trials.

- A count X of successes has a **Poisson distribution** in the **Poisson setting**: the number of successes that occur in two nonoverlapping units of measure are independent; the probability that a success will occur in a unit of measure is the same for all units of equal size and is proportional to the size of the unit; the probability that more than one event occurs in a unit of measure is negligible for very small-sized units. In other words, the events occur one at a time.

- If X has the Poisson distribution with mean μ, then the standard deviation of X is $\sqrt{\mu}$, and the possible values of X are the whole numbers $0, 1, 2, 3$, and so on.

- The **Poisson probability** that X takes any of these values is

$$P(X = k) = \frac{e^{-\mu}\mu^k}{k!} \quad k = 0, 1, 2, 3, ...$$

  Sums of independent Poisson random variables also have the Poisson distribution. For example, in a Poisson model with mean μ per unit of measure, the count of successes in a units is a Poisson random variable with mean $a\mu$.

3 Chapitre 6

3.1 6.1 Estimating with Confidence

- The purpose of a **confidence interval** is to estimate an unknown parameter with an indication of how accurate the estimate is and of how confident we are that the result is correct.
- Any confidence interval has two parts: an interval computed from the data and a confidence level. The interval often has the form

$$\text{estimate} \pm \text{margin of error}$$

- The **confidence level** states the probability that the method will give a correct answer. That is, if you use 95% confidence intervals, in the long run 95% of your intervals will contain the true parameter value. When you apply the method once (that is, to a single sample), you do not know if your interval gave a correct answer (this happens 95% of the time) or not (this happens 5% of the time).
- The **margin of error** for a level C confidence interval for the mean μ of a Normal population with known standard deviation σ, based on an SRS of size n, is given by

$$m = z^* \frac{\sigma}{\sqrt{n}}$$

  Here z^* is obtained from the row labeled z^* at the bottom of **Table D**. The probability is C that a standard Normal random variable takes a value between $-z^*$ and z^*. The confidence interval is

$$\bar{x} \pm m$$

  If the population is not Normal and n is large, the confidence level of this interval is approximately correct.
- Other things being equal, the margin of error of a confidence interval decreases as
  - the confidence level C decreases,
  - the sample size n increases, and
  - the population standard deviation σ decreases.
- The sample size n required to obtain a confidence interval of specified margin of error m for a population mean is

$$n = \left(\frac{z^* \sigma}{m} \right)^2$$

  where z^* is the critical point for the desired level of confidence.
- A specific confidence interval formula is correct only under specific conditions. The most important conditions concern the method used to produce the data. Other factors such as the form of the population distribution may also be important. These conditions should be investigated prior to any calculations.

3.2 6.2 Tests of Significance

- A **test of significance** is intended to assess the evidence provided by data against a **null hypothesis** H_0 in favor of an **alternative hypothesis** H_a.
- The hypotheses are stated in terms of population parameters. Usually, H_0 is a statement that no effect or no difference is present, and H_a says that there is an effect or difference in a specific direction (**one-sided alternative**) or in either direction (**two-sided alternative**).
- The test is based on a **test statistic**. The **P-value** is the probability, computed assuming that H_0 is true, that the test statistic will take a value at least as extreme as that actually observed. Small P-values indicate strong evidence against H_0. Calculating P-values requires knowledge of the sampling distribution of the test statistic when H0 is true.

- If the P-value is as small or smaller than a specified value α, the data are **statistically significant** at significance level α.
- Significance tests for the hypothesis $H_0 : \mu = \mu_0$ concerning the unknown mean μ of a population are based on the z **statistic**:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

The z test **assumes** an SRS of size n, known population standard deviation α, and either a Normal population or a large sample. P-values are computed from the Normal distribution (**Table A**). Fixed α tests use the table of **standard Normal critical values (Table D)**.

3.3 6.3 Use and Abuse of Tests

- P-values are more informative than the reject-or-not result of a level α test. Beware of placing too much weight on traditional values of α, such as $\alpha = 0.05$.
- Very small effects can be highly significant (small P), especially when a test is based on a large sample. A statistically significant effect need not be practically important. Plot the data to display the effect you are seeking, and use confidence intervals to estimate the actual values of parameters.
- On the other hand, lack of significance does not imply that H_0 is true, especially when the test has a low probability of detecting an effect.
- Significance tests are not always valid. Faulty data collection, outliers in the data, and testing a hypothesis on the same data that suggested the hypothesis can invalidate a test. Many tests run at once will probably produce some significant results by chance alone, even if all the null hypotheses are true.

3.4 Power and Inference as a Decision

- The **power** of a significance test measures its ability to detect an alternative hypothesis. The power to detect a specific alternative is calculated as the probability that the test will reject H_0 when that alternative is true. This calculation requires knowledge of the sampling distribution of the test statistic under the alternative hypothesis. Increasing the size of the sample increases the power when the significance level remains fixed.
- An alternative to significance testing regards H_0 and H_a as two statements of equal status that we must decide between. This **decision theory** point of view regards statistical inference in general as giving rules for making decisions in the presence of uncertainty.
- In the case of testing H_0 versus H_a, decision analysis chooses a decision rule on the basis of the probabilities of two types of error. A **Type I error** occurs if H_0 is rejected when it is in fact true. A **Type II error** occurs if H_0 is accepted when in fact H_a is true. ($Power = 1 - TypeII_{error} = 1 - \beta$)
- In a fixed level α significance test, the significance level α is the probability of a Type I error, and the power to detect a specific alternative is 1 minus the probability of a Type II error for that alternative.

4 Chapitre 7: Inference for Means

4.1 7.1 Inference for the Mean of a Population

- Significance tests and confidence intervals for the mean μ of a Normal population are based on the sample mean $\bar{X}$ of an *SRS*. Because of the central limit theorem, the resulting procedures are approximately correct for other population distributions when the sample is large.
- The **standard error** of the sample mean is

$$SE_{\bar{X}} = \frac{s}{\sqrt{n}}$$

- The standardized sample mean, or **one-sample z statistic**,

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

has the $N(0,1)$ distribution. If the standard deviation $\sigma/\sqrt{n}$ of $\bar{x}$ is replaced by the **standard error** $s/\sqrt{n}$, the **one-sample t statistic**

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

has the t **distribution** with $n-1$ degrees of freedom.

- There is a t distribution for every positive **degrees of freedom** k. All are symmetric distributions similar in shape to Normal distributions. The $t(k)$ distribution approaches the $N(0,1)$ distribution as k increases.
- A level C **confidence interval for the mean** μ of a Normal population is

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}}$$

where t^* is the value for the $t(n-1)$ density curve with area C between $-t^*$ and t^*. The quantity

$$t^* \frac{s}{\sqrt{n}}$$

is the **margin of error**.

- Significance tests for $H_0 : \mu = \mu_0$ are based on the t statistic. P-values or fixed significance levels are computed from the $t(n-1)$ distribution.
- A matched pairs analysis is needed when subjects or experimental units are matched in pairs or when there are two measurements on each individual or experimental unit and the question of interest concerns the difference between the two measurements.
- The one-sample procedures are used to analyze **matched pairs** data by first taking the differences within the matched pairs to produce a single sample.
- One-sample **equivalence testing** assesses whether a population mean μ is practically different from a hypothesized mean μ_0. This test requires a threshold δ, which represents the largest difference between μ and μ_0 such that the means are considered equivalent.
- The t procedures are relatively **robust** against non-Normal populations. The t procedures are useful for non-Normal data when $15 \leq n < 40$ unless the data show outliers or strong skewness. When $n \geq 40$, the t procedures can be used even for clearly skewed distributions.

4.2 7.2 Comparing Two Means

- Significance tests and confidence intervals for the difference between the means μ_1 and μ_2 of two Normal populations are based on the difference $\bar{x}_1 - \bar{x}_2$ between the sample means from two independent SRSs. Because of the central limit theorem, the resulting procedures are approximately correct for other population distributions when the sample sizes are large.
- When independent SRSs of sizes n_1 and n_2 are drawn from two Normal populations with parameters μ_1, σ_1 and μ_2, σ_2 the **two-sample z statistic**

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

  has the $N(0,1)$ distribution.
- The **two-sample t statistic**

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

  does not have a t distribution. However, good approximations are available.
- **Conservative inference procedures** for comparing μ_1 and μ_2 are obtained from the two-sample t statistic by using the $t(k)$ distribution with degrees of freedom k equal to the smaller of $n_1 - 1$ and $n_2 - 1$.
- **More accurate probability values** can be obtained by estimating the degrees of freedom from the data. This is the usual procedure for statistical software.
- An approximate level C **confidence interval** for $\mu_1 - \mu_2$ is given by

$$(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

  Here, t^* is the value for the $t(k)$ density curve with area C between $-t^*$ and t^*, where k is computed from the data by software or is the smaller of $n_1 - 1$ and $n_2 - 1$. The quantity

$$t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

  is the **margin of error**.
- Significance tests for $H_0 : \mu_1 - \mu_2 = \Delta_0$ use the **two-sample t statistic**

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

  The P-value is approximated using the $t(k)$ distribution where k is estimated from the data using software or is the smaller of $n_1 - 1$ and $n_2 - 1$.
- The guidelines for practical use of two-sample t procedures are similar to those for one-sample t procedures. Equal sample sizes are recommended.
- If we can assume that the two populations have equal variances, **pooled two-sample t procedures** can be used. These are based on the **pooled estimator**

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

  of the unknown common variance and the $t(n_1 + n_2 - 2)$ distribution. We do not recommend this procedure for regular use.

4.3 7.3 Additional Topics on Inference

- The **sample size** required to obtain a confidence interval with an expected margin of error no larger than m for a population mean satisfies the constraint

$$m \geq t^* s^* / \sqrt{n}$$

where t^* is the critical value for the desired level of confidence with $n-1$ degrees of freedom, and s^* is the guessed value for the population standard deviation.

- The sample sizes necessary for a two-sample confidence interval can be obtained using a similar constraint, but guesses of both standard deviations and an estimate for the degrees of freedom are required. We suggest using the smaller of $n_1 - 1$ and $n_2 - 1$ for degrees of freedom.

- The **power** of the one-sample t test can be calculated like that of the z test, using an approximate value for both σ and s.

- The **power** of the two-sample t test is found by first finding the critical value for the significance test, the degrees of freedom, and the **noncentrality parameter** for the alternative of interest. These are used to calculate the power from a **noncentral t distribution**. A Normal approximation works quite well. Calculating margins of error for various study designs and conditions is an alternative procedure for evaluating designs.

- The **sign test** is a **distribution-free test** because it uses probability calculations that are correct for a wide range of population distributions.

- The sign test for "no treatment effect" in matched pairs counts the number of positive differences. The P-value is computed from the $B(n, 1/2)$ distribution, where n is the number of non-0 differences. The sign test is less powerful than the t test in cases where use of the t test is justified.

5 Chapitre 8: Inference for Proportions

5.1 Inference for a Single Proportion

- Inference about a population proportion p from an SRS of size n is based on the **sample proportion** $\hat{p} = X/n$. When n is large, $\hat{p}$ has approximately the Normal distribution with mean p and standard deviation $\sqrt{p(1-p)/n}$.
- For large samples, the **margin of error for confidence level** C is

$$m = z^* SE_{\hat{p}}$$

where the critical value z^* is the value for the standard Normal density curve with area C between $-z^*$ and z^*, and the **standard error** of $\hat{p}$ is

$$SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- The **level** C **large-sample confidence interval** is

$$\hat{p} \pm m$$

We recommend using this interval for $90\%, 95\%$, and 99% confidence whenever the number of successes and the number of failures are both at least 10. When sample sizes are smaller, alternative procedures such as the plus four estimate of the population proportion are recommended.

- The **sample size** required to obtain a confidence interval of approximate margin of error m for a proportion is found from

$$n = \left(\frac{z^*}{m}\right)^2 p^*(1-p^*)$$

where p^* is a guessed value for the proportion and z^* is the standard Normal critical value for the desired level of confidence. To ensure that the margin of error of the interval is less than or equal to m no matter what $\hat{p}$ may be, use

$$n = \frac{1}{4}\left(\frac{z^*}{m}\right)^2$$

- Tests of $H_0 : p = p_0$ are based on the z **statistic**

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

with P-values calculated from the $N(0,1)$ distribution. Use this procedure when the expected number of successes, np_0, and the expected number of failures, $n(1-p_0)$, are both greater than 10.

- Software can be used to determine the sample sizes for significance tests.

5.2 8.2 Comparing Two Proportions

- The **large-sample estimate of the difference in two population proportions** is

$$D = \hat{p}_1 - \hat{p}_2$$

where $\hat{p}_1$ and $\hat{p}_2$ are the sample proportions:

$$\hat{p}_1 = \frac{X_1}{n_1} and \hat{p}_2 = \frac{X_2}{n_2}$$

- The **standard error of the difference** D is

$$SE_D = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

- The **margin of error for confidence level** C is

$$m = z^* SE_D$$

where z^* is the value for the standard Normal density curve with area C between $-z^*$ and z^*. The **large-sample level** C **confidence interval** is

$$D \pm m$$

We recommend using this interval for $90\%, 95\%$, or 99% confidence when the number of successes and the number of failures in both samples are all at least 10. When sample sizes are smaller, alternative procedures such as the plus four estimate of the difference in two population proportions are recommended.

- Significance tests of $H_0 : p_1 = p_2$ use the z **statistic**

$$z = \frac{\hat{p}_1 - \hat{p}_2}{SE_{D_p}}$$

with P-values from the $N(0, 1)$ distribution. In this statistic,

$$SE_{D_p} = \sqrt{\hat{p}(1-\hat{p})\Big(\frac{1}{n_1} + \frac{1}{n_2}\Big)}$$

and $\hat{p}$ is the **pooled estimate** of the common value of p_1 and p_2:

$$\hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$$

Use this test when the number of successes and the number of failures in each of the samples are at least 5.

- **Relative risk** is the ratio of two sample proportions:

$$RR = \frac{\hat{p}_1}{\hat{p}_2}$$

Confidence intervals for relative risk are often used to summarize the comparison of two proportions.

6 Chapitre 9: Inference for Categorical Data

6.1 Inference for Two-Way Tables

- Soit $Z_1, Z_2, ..., Z_p$ des variables aléatoires i.i.d. normale standard et soit $X = \sum_{i=1}^{p} Z_i^2$. Alors X suit une distribution **chi-deux** à p degrés de liberté, dénotée par χ_p^2.
- The **null hypothesis** for $r \times c$ tables of count data is that there is no relationship between the row variable and the column variable.
- **Expected cell counts** under the null hypothesis are computed using the formula

$$\text{expected count} = \frac{\text{row total} \times \text{column total}}{n}$$

- The null hypothesis is tested by the **chi-square statistic**, which compares the observed counts with the expected counts:

$$X^2 = \sum \frac{\big(\text{observed} - \text{expected}\big)^2}{\text{expected}}$$

  Under the null hypothesis, X^2 has approximately the χ^2 distribution with $(r-1)(c-1)$ degrees of freedom. The P-value for the test is

$$P\Big(\chi^2 \geq X^2\Big)$$

  where χ^2 is a random variable having the $\chi^2(df)$ distribution with $df = (r-1)(c-1)$.
- The chi-square approximation is adequate for practical use when the average expected cell count is 5 or greater and all individual expected counts are 1 or greater, except in the case of 2×2 tables. All four expected counts in a 2×2 table should be 5 or greater.
- For two-way tables, we first compute percents or proportions that describe the relationship of interest. Then, we compute expected counts, the X^2 statistic, and the P-value.
- Two different models for generating $r \times c$ tables lead to the chi-square test. In the first model, independent simple random samples (SRSs) are drawn from each of c populations, and each observation is classified according to a categorical variable with r possible values. The null hypothesis is that the distributions of the row categorical variable are the same for all c populations. In the second model, a single SRS is drawn from a population, and observations are classified according to two categorical variables having r and c possible values. In this model, H_0 states that the row and column variables are independent.

6.2 9.2 Goodness

- The **chi-square goodness-of-fit test** is used to compare the sample distribution of a categorical variable from a population with a hypothesized distribution. The data for n observations with k possible outcomes are summarized as observed counts,$n_1, n_2, ..., n_k$, in k cells. The **null hypothesis** specifies probabilities $p_1, p_2, ..., p_k$ for the possible outcomes.
- The analysis of these data is similar to the analyses of two-way tables discussed in Section 9.1. For each cell, the **expected count** is determined by multiplying the total number of observations n by the specified probabilitypi. The null hypothesis is tested by the usual **chi-square statistic**, which compares the observed counts, ni, with the expected counts. Under the null hypothesis, X^2 has approximately the χ^2 distribution with $df = k - 1$.

7 Chapitre 10: Inference for Regression

7.1 10.1 Simple Linear Regression

- The statistical model for **simple linear regression** assumes that the means of the response variable y fall on a line when plotted against x, with the observed y's varying Normally about these means. For n observations, this model can be written

$$y_i = \beta_0 + \beta_1 \mathsf{x}_i + \epsilon_i$$

where $i = 1, 2, ..., n$, and the ϵ_i are assumed to be independent and Normally distributed with mean 0 and standard deviation σ. Here $\beta_0 + \beta_1 \mathsf{x}_i$ is the mean response when $\mathsf{x} = \mathsf{x}_i$. The **parameters** of the model are β_0, β_1, and σ.

- The **population regression line** intercept and slope, β_0 and β_1, are estimated by the intercept and slope of the **least-squares regression line**, b_0 and b_1. The **model standard deviation** σ is estimated by

$$s = \sqrt{\frac{\sum e_i^2}{n-2}}$$

where the e_i are the **residuals**

$$e_i = y_i - \hat{y}_i$$

- Prior to inference, always examine the residuals for Normality, constant variance, and any other remaining patterns in the data. **Plots of the residuals** both against the case number and against the explanatory variable are commonly part of this examination. Scatterplot smoothers are helpful in detecting patterns in these plots.

- A **level** C **confidence interval for** β_1 is

$$b_1 \pm t^* SE_{b_1}$$

where t^* is the value for the $t(n-2)$ density curve with area C between $-t^*$ and t^*.

- The **test of the hypothesis** $H_0 : \beta_1 = 0$ is based on the t **statistic**

$$t = \frac{b_1}{SE_{b_1}}$$

and the $t(n-2)$ distribution. This tests whether there is a straight-line relationship between y and x. There are similar formulas for confidence intervals and tests for β_0, but these are meaningful only in special cases.

- The **estimated mean response** for the subpopulation corresponding to the value x^* of the explanatory variable is

$$\widehat{\mu}_y = b_0 + b_1 x^*$$

- A **level** C **confidence interval for the mean response** is

$$\widehat{\mu_y} \pm t^* SE_{\widehat{\mu}}$$

where t^* is the value for the $t(n-2)$ density curve with area C between $-t^*$ and t^*.

- The **estimated value of the response variable** y for a future observation from the subpopulation corresponding to the value x^* of the explanatory variable is

$$\hat{y} = b_0 + b_1 x^*$$

- A **level** C **prediction interval** for the estimated response is

$$\hat{y} \pm t^* SE_{\hat{y}}$$

where t^* is the value for the $t(n-2)$ density curve with area C between $-t^*$ and t^*. The standard error for the prediction interval is larger than the confidence interval because it also includes the variability of the future observation around its subpopulation mean.

- Sometimes, a **transformation** of one or both of the variables can make their relationship linear. However, these transformations can harm the assumptions of Normality and constant variance, so it is important to examine the residuals.

7.2 10.2 More Detail about Simple Linear Regression

- The **ANOVA table** for a linear regression gives the degrees of freedom, sum of squares, and mean squares for the model, error, and total sources of variation. The **ANOVA** F **statistic** is the ratio MSM/MSE. Under $H_0 : \beta_1 = 0$, this statistic has an $F(1, n-2)$ distribution and is used to test H_0 versus the two-sided alternative.

- The **square of the sample correlation** can be expressed as

$$r^2 = \frac{SSM}{SST}$$

and is interpreted as the proportion of the variability in the response variable y that is explained by the explanatory variable x in the linear regression.

- The **standard errors for** b_0 **and** b_1 are

$$SE_{b_0} = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}}$$
$$SE_{b_1} = \frac{s}{\sqrt{\sum(x_i - \bar{x})^2}}$$

- The **standard error that we use for a confidence interval** for the estimated mean response for the subpopulation corresponding to the value x^* of the explanatory variable is

$$SE_{\hat{\mu}} = s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

- The **standard error that we use for a prediction interval** for a future observation from the subpopulation corresponding to the value x^* of the explanatory variable is

$$SE_{\hat{y}} = s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

- When the variables y and x are jointly Normal, the sample correlation is an estimate of the **population correlation** ρ. The test of $H_0 : \rho = 0$ is based on the t **statistic**

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

which has a $t(n-2)$ distribution under H_0. This test statistic is numerically identical to the t statistic used to test $H_0 : \beta_1 = 0$.

Index

V