MATH 205: Statistical methods

December 8th, 2021

Review

Announcements

- Final exam: next Monday (12/13) at 3:30pm.
- Closed-book. You are allowed to bring a one-sided hand-written A4-sized note to the exam.
- You can use calculators (and you should have one).
- Course evaluation

Expected value: discrete variables

Definition

Given a discrete random variable X which takes values in the set $\mathcal{D}$ and which has probability distribution P, we define the expected value of X as

$$\mathbb{E}[X] = \sum_{x \in \mathcal{D}} xP(X = x)$$

This is sometimes written $\mathbb{E}_P[X]$, to clarify which distribution one has in mind.

Expected value: continuous variables

Definition

Given a discrete random variable X which takes values in the set $\mathcal{D}$ and which has probability density function $p(x)$, we define the expected value of X as

$$\mathbb{E}[X] = \int_{\mathcal{D}} xp(x)\ dx$$

This is sometimes written $\mathbb{E}_P[X]$, to clarify which distribution one has in mind.

Mean and variance

Definition

- The mean or expected value of a random variable X is

$$\mathbb{E}[X]$$

- The variance of a random variable X is

$$var[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

- The standard deviation of a random variable X is defined as

$$std(X) = \sqrt{var(X)}$$

Expected value: discrete variables

Definition

Assume we have a function f that maps a discrete random variable X into a set of numbers D_f. Then f(X) is a discrete random variable, too, which we write F. The expected value of this random variable is written

$$\mathbb{E}[f(X)] = \sum_{x \in \mathcal{D}} f(x) P(X = x)$$

which is sometimes referred to as "the expectation of f". The process of computing an expected value is sometimes referred to as "taking expectations".
This is sometimes written $\mathbb{E}[f]$, or $\mathbb{E}_P[f]$ or $\mathbb{E}_{P(X)}[f]$.

Expected value: continuous variables

Definition

Assume we have a function f that maps a discrete random variable X into a set of numbers D_f. Then f(X) is a continuous random variable, too, which we write F. The expected value of this random variable is written

$$\mathbb{E}[f(X)] = \int_{\mathcal{D}} f(x)p(x)\ dx$$

which is sometimes referred to as "the expectation of f". The process of computing an expected value is sometimes referred to as "taking expectations".
This is sometimes written $\mathbb{E}[f]$, or $\mathbb{E}_P[f]$ or $\mathbb{E}_{P(X)}[f]$.

Linear combination of random variables

Theorem

Let $X_1, X_2, \ldots, X_n$ *be independent random variables (with possibly different means and/or variances). Define*

$$T = a_1X_1 + a_2X_2 + \ldots + a_nX_n$$

then the mean and the standard deviation of T *can be computed by*

- $E(T) = a_1E(X_1) + a_2E(X_2) + \ldots + a_nE(X_n)$
- $Var(T) = a_1^2Var(X_1) + a_2^2Var(X_2) + \ldots + a_n^2Var(X_n)$

Linear combination of normal random variables

Theorem

Let $X_1, X_2, \ldots, X_n$ be independent normal random variables (with possibly different means and/or variances). Then

$$T = a_1 X_1 + a_2 X_2 + \ldots a_n X_n$$

also follows the normal distribution with

- $E(T) = a_1 E(X_1) + a_2 E(X_2) + \ldots + a_n E(X_n)$
- $Var(T) = a_1^2 Var(X_1) + a_2^2 Var(X_2) + \ldots + a_n^2 Var(X_n)$

Basic properties of probability

Useful Facts 3.1 (Basic Properties of the Probability Events)
We have

- The probability of every event is between zero and one; in equations

$$0 \leq P(\mathcal{A}) \leq 1$$

  for any event $\mathcal{A}$.
- Every experiment has an outcome; in equations,

$$P(\Omega) = 1.$$

- The probability of disjoint events is additive; writing this in equations requires some notation. Assume that we have a collection of events $\mathcal{A}_i$, indexed by i. We require that these have the property $\mathcal{A}_i \cap \mathcal{A}_j = \varnothing$ when $i \neq j$. This means that there is no outcome that appears in more than one $\mathcal{A}_i$. In turn, if we interpret probability as relative frequency, we must have that

$$P(\cup_i \mathcal{A}_i) = \sum_i P(\mathcal{A}_i)$$

Advanced properties of probability

Useful Facts 3.2 (Properties of the Probability of Events)

- $P(\mathcal{A}^c) = 1 - P(\mathcal{A})$
- $P(\varnothing) = 0$
- $P(\mathcal{A} - \mathcal{B}) = P(\mathcal{A}) - P(\mathcal{A} \cap \mathcal{B})$
- $P(\mathcal{A} \cup \mathcal{B}) = P(\mathcal{A}) + P(\mathcal{B}) - P(\mathcal{A} \cap \mathcal{B})$

- If $A \subset B$, then $P(A) \leq P(B)$.
- For any events A, B

$$P(A) = P(A \cap B) + P(A \cap B^c)$$

Independence

Definition

Two events A and B are independent if and only if

$$P(A \cap B) = P(A)P(B)$$

Conditional probability

Definition

Let $P(A) > 0$, the conditional probability of B given A, denoted by $P(B|A)$, is

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

Properties of Conditional probability

- Law of multiplication

$$P(B \cap A) = P(B|A)P(A)$$

- **Bayes' rule**

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

- Law of total probability

$$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c)$$

Correlation coefficient

Definition 2.1 (Correlation Coefficient) Assume we have N data items which are 2-vectors $(x_1, y_1), \ldots, (x_N, y_N)$, where $N > 1$. These could be obtained, for example, by extracting components from larger vectors. We compute the correlation coefficient by first normalizing the x and y coordinates to obtain $\hat{x}_i = \frac{(x_i - \mathsf{mean}(\{x\}))}{\mathsf{std}(x)}$, $\hat{y}_i = \frac{(y_i - \mathsf{mean}(\{y\}))}{\mathsf{std}(y)}$. The correlation coefficient is the mean value of $\hat{x}\hat{y}$, and can be computed as:

$$\mathsf{corr}(\{(x, y)\}) = \frac{\sum_i \hat{x}_i \hat{y}_i}{N}$$

Correlation coefficient: properties

Useful Facts 2.1 (Properties of the Correlation Coefficient)

- The correlation coefficient is symmetric (it doesn't depend on the order of its arguments), so

$$\mathsf{corr}\left(\{(x, y)\}\right) = \mathsf{corr}\left(\{(y, x)\}\right)$$

- The value of the correlation coefficient is not changed by translating the data. Scaling the data can change the sign, but not the absolute value. For constants $a \neq 0$, b, $c \neq 0$, d we have

$$\mathsf{corr}\left(\{(ax + b, cx + d)\}\right) = \text{sign}(ab)\mathsf{corr}\left(\{(x, y)\}\right)$$

- If $\hat{y}$ tends to be large (resp. small) for large (resp. small) values of $\hat{x}$, then the correlation coefficient will be positive.
- If $\hat{y}$ tends to be small (resp. large) for large (resp. small) values of $\hat{x}$, then the correlation coefficient will be negative.
- If $\hat{y}$ doesn't depend on $\hat{x}$, then the correlation coefficient is zero (or close to zero).
- The largest possible value is 1, which happens when $\hat{x} = \hat{y}$.
- The smallest possible value is -1, which happens when $\hat{x} = -\hat{y}$.

Using correlation to predict

Procedure 2.1 (Predicting a Value Using Correlation) Assume we have N data items which are 2-vectors $(x_1, y_1), \ldots, (x_N, y_N)$, where $N > 1$. These could be obtained, for example, by extracting components from larger vectors. Assume we have an x value x_0 for which we want to give the best prediction of a y value, based on this data. The following procedure will produce a prediction:

- Transform the data set into standard coordinates, to get

$$\hat{x}_i = \frac{1}{\mathsf{std}\,(x)}(x_i - \mathsf{mean}\,(\{x\}))$$

$$\hat{y}_i = \frac{1}{\mathsf{std}\,(y)}(y_i - \mathsf{mean}\,(\{y\}))$$

$$\hat{x}_0 = \frac{1}{\mathsf{std}\,(x)}(x_0 - \mathsf{mean}\,(\{x\})).$$

- Compute the correlation

$$r = \mathsf{corr}\,(\{(x, y)\}) = \mathsf{mean}\,(\{\hat{x}\hat{y}\}).$$

- Predict $\hat{y}_0 = r\hat{x}_0$.
- Transform this prediction into the original coordinate system, to get

$$y_0 = \mathsf{std}\,(y) r\hat{x}_0 + \mathsf{mean}\,(\{y\})$$

Test about a population mean

- Null hypothesis

$$H_0 : \mu = \mu_0$$

- The alternative hypothesis will be either:
  - $H_a : \mu > \mu_0$
  - $H_a : \mu < \mu_0$
  - $H_a : \mu \neq \mu_0$

Note: μ_0 here denotes a constant, and μ denotes the population mean (unknown)

We use the test statistic:

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

P-values for z-tests

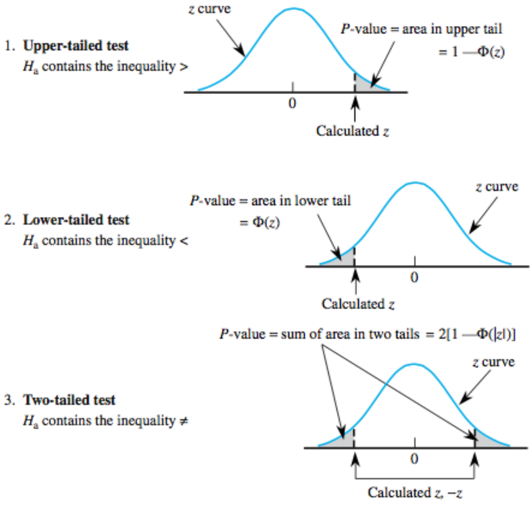

Figure 9.7 Determination of the P-value for a z test

Practice problem

Problem

The target thickness for silicon wafers used in a certain type of integrated circuit is 245 μm. A sample of 50 wafers is obtained and the thickness of each one is determined, resulting in a sample mean thickness of 246.18 μm and a sample standard deviation of 3.60 μm.
At significant level $\alpha = 0.01$, does this data suggest that true average wafer thickness is something other than the target value?

P-values for z-tests

1. Parameter of interest: μ = true average wafer thickness
2. Null hypothesis: H_0: $\mu = 245$
3. Alternative hypothesis: H_a: $\mu \neq 245$
4. Formula for test statistic value: $z = \dfrac{\bar{x} - 245}{s/\sqrt{n}}$
5. Calculation of test statistic value: $z = \dfrac{246.18 - 245}{3.60/\sqrt{50}} = 2.32$
6. Determination of P-value: Because the test is two-tailed,

$$P\text{-value} = 2[1 - \Phi(2.32)] = .0204$$

7. Conclusion: Using a significance level of .01, H_0 would not be rejected since .0204 > .01. At this significance level, there is insufficient evidence to conclude that true average thickness differs from the target value.

Testing the difference between two population means

Assume that we want to test the null hypothesis $H_0 : \mu_1 - \mu_2 = \Delta_0$ against each of the following alternative hypothesis

(a) $H_a : \mu_1 - \mu_2 > \Delta_0$

(b) $H_a : \mu_1 - \mu_2 < \Delta_0$

(c) $H_a : \mu_1 - \mu_2 \neq \Delta_0$

We use the test statistic:

$$z = \frac{(\bar{x} - \bar{y}) - \Delta_0}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}}.$$

and derive the p-value in the same way as the one-sample tests.