CS 397: Topics in Computer Science–Probability & Statistics

Dr. Francis Parisi

Pace University

Early Summer 2019

Probability & Statistics for Computer Science

- Expected Value of a Random Variable
- Moments
- Chebyshev's Theorem
- Moment-Generating Functions
- Product Moments
- Moments of Linear Combinations of Random Variables
- Conditional Expectations

Probability & Statistics for Computer Science

Expected Value of a Random Variable

- The **expected value** of a random variable is

  $$\mathbb{E}\,X = \sum_x xf(x)$$

  for discrete random variables and

  $$\mathbb{E}\,X = \int_{-\infty}^{\infty} xf(x)dx$$

  for continuous random variables
- Expectation is a linear operator

  $$\mathbb{E}[aX + b] = a\,\mathbb{E}\,X + b$$

Probability & Statistics for Computer Science

Moments

- We define the **moments** of a random variable as

$$\mu_r' = \mathbb{E}[X^r]$$

- The first moment is the mean
- We define moments around the mean as

$$\mu_r = \mathbb{E}[(X-\mu)^r]$$

  and we saw earlier that when $r = 2$ this becomes the $\mathrm{Var}\left(X\right)$

- The variance is also linear as we'll talk more about this later $\mathrm{Var}\left(aX+b\right) = a^2\mathrm{Var}\left(X\right) = a^2\sigma^2$ What happened to b?

- Chebyshev's Theorem gives us a lower bound on the the probability that a given random variable will take on a value within k standard deviations of the mean

Theorem

If μ and σ are the mean and the standard deviation of a random variable X, then for any positive constant k the probability is at least $1 - \frac{1}{k^2}$ that X will take on a value within k standard deviations of the mean

Symbolically we have

$$P\left[|x - \mu| \leq k\sigma\right] \geq 1 - \frac{1}{k^2}$$

Probability & Statistics for Computer Science

Moment Generating Functions

- While we can find any moment by the direct integration or summation, alternatively we can use the **moment generating function**
- Moment Generating Function of the random variable X

$$M_X(t) = \mathbb{E}[e^{tX}]$$

- $M_X(t) = \sum_x e^{tx} f(x)$ if X is discrete
- $M_X(t) = \int_{-\infty}^{\infty} e^{tx} f(x) dx$ if X is continuous

Probability & Statistics for Computer Science

We can easily find the first few moments using the derivative of the mgf plugging in $t = 0$

$$\frac{d^r M_X(t)}{dt^r}\Big|_{t=0} = \mu_r'$$

Example

Random variable X has pmf $f(x) = \frac{1}{8}\binom{3}{x}, x = 0, 1, 2, 3$. Find the first two moments.

Probability & Statistics for Computer Science

Point Estimation

Probability & Statistics for Computer Science

Definition

A statistic $\hat{\Theta}$ is said to be an unbiased estimator of the parameter θ if

$$\mathbb{E}[\hat{\Theta}] = \theta$$

Example

If X has the binomial distribution with the parameters n and θ, show that the sample proportion, X/n, is an unbiased estimator of θ.

Solution

Since $\mathbb{E}[X] = n\theta$, then

$$\mathbb{E}\left[\frac{X}{n}\right] = \frac{1}{n}\,\mathbb{E}[X] = \frac{1}{n}n\theta = \theta$$

Definition

The estimator for the parameter θ of a given distribution that has the smallest variance of all unbiased estimators for θ is called the **minimum variance unbiased estimator**, or the best unbiased estimator for θ. The estimator with the smallest variance is an **efficient** estimator.

Definition

The statistic $\hat{\Theta}$ is a consistent estimator of the parameter θ of a given distribution if and only if for each $c > 0$

$$\lim_{n \to \infty} P\left[|\hat{\Theta} - \theta| < c\right] = 1$$

Probability & Statistics for Computer Science

Definition

The statistic $\hat{\Theta}$ is a **sufficient estimator** of the parameter θ of a given distribution if and only if for each value of $\hat{\Theta}$ the conditional probability distribution or density of the random sample $X_1, X_2, \ldots, X_n$, given $\hat{\Theta} = \theta$, is independent of θ.

A sufficient estimator uses all the information in the sample

Probability & Statistics for Computer Science

Hypothesis Testing

- Hypothesis testing is a frequently used method of inference
- We start by stating the **null hypothesis**, H_0, and the **alternative hypothesis**, H_1 (sometimes denoted by H_a)
- For example $H_0 : \mu = 0$, and $H_1 : \mu \neq 0$ this is a two-sided or two-tailed test
- Or $H_1 : \mu < 0$ (one-sided or one-tailed test, left tail), $H_1 : \mu > 0$ (right tail test)
- There are several steps to complete the hypothesis test

Probability & Statistics for Computer Science

Steps in hypothesis testing:

Step 1: State the null and alternative hypotheses, H_0, and H_1, determine if this is a one-tailed or two-tailed test

Step 2: State α, the level of significance

Step 3: Calculate the test statistic from the sample mean and variance

$$t = \frac{\overline{X} - \mu}{\sqrt{s^2/n}}$$

Step 4: Compare the calculated test statistic to the critical value

Step 5: Accept H_0 if t is in the acceptance region, or reject H_0 if t is in the critical or rejection region

Probability & Statistics for Computer Science

- We set the level of significance α *before* we do the test
- Often $\alpha = 0.05$ and represents the **size** of the critical region
- If the test statistic t falls in the critical region, we reject H_0, otherwise we do not reject H_0
- If the data come from a normal distribution or the sample size is large the test statistic follows a standard normal distribution rather than a t-distribution

$$Z = \frac{\overline{X} - \mu}{\sqrt{s^2/n}}$$

- There's always a risk of making a wrong decision and these are known as **Type I** and **Type II** errors

Definition

1. Rejecting H_0 when it is true is called a Iype I error. The probability of committing a type I error is denoted by α
2. Not rejecting H_0 when it is false is called a Type II error. The probability of committing a type II error is denoted by β

Probability & Statistics for Computer Science

	Reality	
Decision	H_0 is True	H_0 is NOT True
Reject H_0	**Type I** Error	Correct Decision
DO NOT Reject H_0	Correct Decision	**Type II** Error

Example

Suppose we sample 26 water bottles for some known chemical. The mean level of the chemical is believed to be 0 mg/dl. We measure each sample and find a sample mean $\overline{X}$ of 1.9 mg/dl of the chemical, with a standard deviation, S equal to 4.2122. Is there sufficient evidence at a $\alpha = 0.05$ level of significance to say the average amount of chemical in the water is not zero?

Probability & Statistics for Computer Science

We have $H_0 : \mu = 0$, $H_1 : \mu \neq 0$, $\alpha = 0.05$, $\overline{X} = 1.9$, $S = 4.2122$, and $n = 26$

therefore,

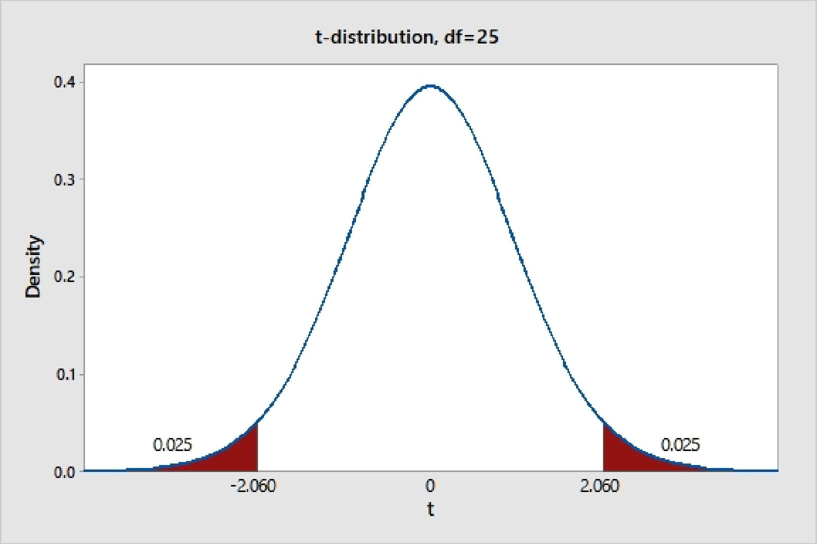
$$t = \frac{1.9 - 0}{4.2122/\sqrt{26}} = 2.3$$

We compare this t to the critical value for $t_{\alpha/2,n-1}$

We find that $t_{\alpha/2,n-1} = 2.06$, thus we reject H_0 since $t > t_{critical}$

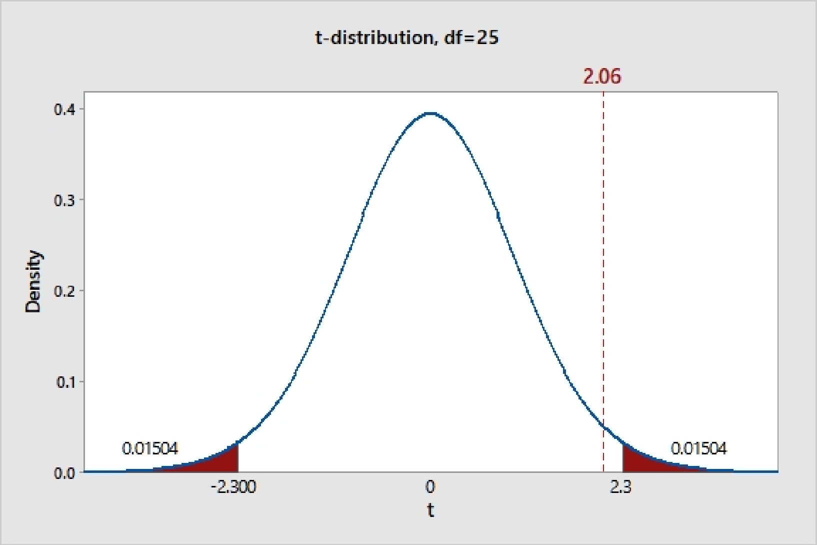
Probability & Statistics for Computer Science

What does the critical value mean?



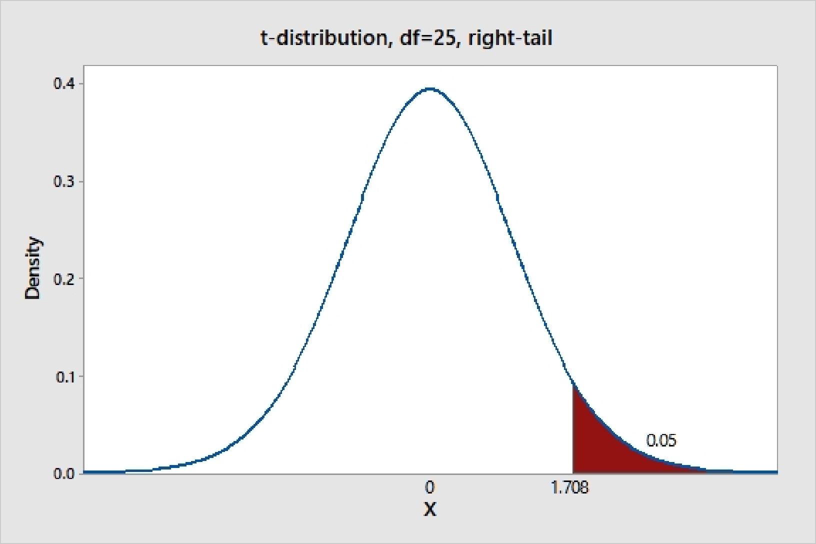
Probability & Statistics for Computer Science

p-values



Probability & Statistics for Computer Science

For a one-tailed test we have



Probability & Statistics for Computer Science

p-values

- The tail probability associated with the t-statistic from our hypothesis test is called the p-value
- If we consider α as our theoretical level of significance then the p-value is the *observed* level of significance
- For a one-tailed test the p-value is the tail probability
- For a two-tailed test it is twice the tail probability
- We *reject the null hypothesis* when $p < \alpha$

Probability & Statistics for Computer Science

Exercise

Suppose that 100 high-performance tires made by a certain manufacturer lasted on the average 21,819 miles with a standard deviation of 1,295 miles. Test the null hypothesis $\mu = 22,000$ miles against the alternative hypothesis $\mu < 22,000$ miles at the 0.05 level of significance.

Tests Concerning Differences Between Means

Probability & Statistics for Computer Science

Tests Concerning Differences Between Means

- Two independent random samples of sizes n_1 and n_2 respectively
- Drawn from two populations with means μ_1 and μ_2, and variance σ_1^2 and σ_2^2
- The test statistic becomes

$$z = \frac{(\overline{x}_1 - \overline{x}_2) - \delta}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

Probability & Statistics for Computer Science

Example

An experiment is performed to determine whether the average nicotine content of one kind of cigarette exceeds that of another kind by 0.20 milligram. If $n_1 = 50$ cigarettes of the first kind had an average nicotine content of $\bar{x}_1 = 2.61$ milligrams with a standard deviation of $s_1 = 0.12$ milligram, whereas $n_2 = 40$ cigarettes of the other kind had an average nicotine content of $\bar{x}_2 = 2.38$ milligrams with a standard deviation of $s_2 = 0.14$ milligram, test the null hypothesis $\mu_1 - \mu_2 = 0.20$ against the alternative hypothesis $\mu_1 - \mu_2 \neq 0.20$ at the 0.05 level of significance. Base the decision on the p-value corresponding to the value of the appropriate test statistic.

Probability & Statistics for Computer Science

- When n_1 and n_2 are small and σ_1 and σ_2 are unknown we need to change our test a little
- For independent samples drawn from two normal populations with means μ_1 and μ_2, and the same *unknown* variance σ^2
- We use the Two-Sample t-test

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - \delta}{s_p \sqrt{1/n_1 + 1/n_2}}$$

with

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

Probability & Statistics for Computer Science

Example

In the comparison of two kinds of paint, a consumer testing service finds that four 1-gallon cans of one brand cover on the average 546 square feet with a standard deviation of 31 square feet, whereas four 1-gallon cans of another brand cover on the average 492 square feet with a standard deviation of 26 square feet. Assuming that the two populations sampled are normal and have equal variances, test the null hypothesis $\mu_1 - \mu_2 = 0$ against the alternative hypothesis $\mu_1 - \mu_2 > 0$ at the 0.05 level of significance.

Probability & Statistics for Computer Science

Exercise

An experiment was performed to compare the abrasive wear of two different laminated materials. Twelve pieces of material 1 were tested by exposing each piece to a machine measuring wear. Ten pieces of material 2 were similarly tested. In each case, the depth of wear was observed. The samples of material 1 gave an average (coded) wear of 85 units with a sample standard deviation of 4, while the samples of material 2 gave an average of 81 with a sample standard deviation of 5. Can we conclude at the 0.05 level of significance that the abrasive wear of material 1 exceeds that of material 2 by more than 2 units? Assume the populations to be approximately normal with equal variances.

Probability & Statistics for Computer Science

Testing Variances

Probability & Statistics for Computer Science

- We can compare a sample mean to a population mean, or a hypothesized mean
- We can analyze the difference between means
- We now consider testing the variance against a constant or comparing two variances

Probability & Statistics for Computer Science

- When we test for the difference in variances we find that our test statistic is the ratio of two variances
- Under the assumption that the sample comes from a normal population the test statistic for s^2 vs. σ_0^2

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

follows a χ^2 distribution

Probability & Statistics for Computer Science

Example

Suppose that the uniformity of the thickness of a part used in a semiconductor is critical and that measurements of the thickness of a random sample of 18 such parts have the variance $s^2 = 0.68$, where the measurements are in thousandths of an inch. The process is considered to be under control if the variation of the thicknesses is given by a variance not greater than 0.36. Assuming that the measurements constitute a random sample from a normal population, test the null hypothesis $\sigma^2 = 0.36$ against the alternative hypothesis $\sigma^2 > 0.36$ at the 0.05 level of significance.

Probability & Statistics for Computer Science

- Rather than testing a sample variance against a fixed or hypothesized variance, what if we test the equality between two *sample* variances?
- In the two-sample t-test above we assumed the two populations had equal variances, how can we test that assumption?

Probability & Statistics for Computer Science

- Independent random samples from two normal populations with variances σ_1^2 and σ_2^2
- To test $H_0 : \sigma_1^2 = \sigma_2^2$ against an alternative hypothesis, the statistic

$$F = s_1^2/s_2^2$$

  follows an F-distribution

Example

In comparing the variability of the tensile strength of two kinds of structural steel, an experiment yielded the following results: $n_1 = 13, s_1^2 = 19.2, n_2 = 16,$ and $s_2^2 = 3.5,$ where the units of measurement are 1,000 pounds per square inch. Assuming that the measurements constitute independent random samples from two normal populations, test the null hypothesis $\sigma_1^2 = \sigma_2^2$ against the alternative $\sigma_1^2 \neq \sigma_2^2$ at the 0.02 level of significance.

Probability & Statistics for Computer Science

Testing Proportions

- When our interest is in the number of outcomes of a particular type we are dealing with proportions
- Suppose we ask n Pace students if they prefer cafeteria food or home cooking, and x students say cafeteria food, then

  $$\hat{p} = \frac{x}{n}$$

  is the proportion of students who prefer cafeteria food
- We can test hypotheses concerning proportions just like we did with means

Probability & Statistics for Computer Science

- For proportions, $\hat{p}$ is the proportion estimate and $\frac{p(1-p)}{n}$ is the variance
- The test statistic is

$$Z = \frac{x - np_0}{\sqrt{np_0(1-p_0)}} = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

Probability & Statistics for Computer Science

Example

A commonly prescribed drug for relieving nervous tension is believed to be only 60% effective. Experimental results with a new drug administered to a random sample of 100 adults who were suffering from nervous tension show that 70 received relief. Is this sufficient evidence to conclude that the new drug is superior to the one commonly prescribed? Use a 0.05 level of significance.

Probability & Statistics for Computer Science

- As in the case of testing the difference in means, we can test for a difference in proportions
- The test statistics becomes

$$Z = \frac{\hat{p_1} - \hat{p_2}}{\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}}$$

Probability & Statistics for Computer Science

Test of Independence Using χ^2

Probability & Statistics for Computer Science

- Let's look a test to determine if a population has a specified theoretical distribution
- The test is based on how good a fit we have between the frequency of occurrence of observations in an observed sample and the expected frequencies obtained from the hypothesized distribution

Probability & Statistics for Computer Science

Definition

A table having r rows and c columns where each row represents c values of a non-numerical variable and each column represents r values of a different nonnumerical variable is called a **contingency table.** In such a table, the entries are count data (positive integers) and both the row and the column totals are left to chance. Such a table is assembled for the*purpose of testing whether the row variable and the column variable are independent*.

Probability & Statistics for Computer Science

Let’s say we want to study the relationship, if any, of the I.Q.’s of employees who have gone through a large company’s job-training program and their subsequent performance on the job:

		Performance			
		Poor	*Fair*	*Good*	
	Below Average	67	64	25	156
I.Q.	*Average*	42	76	56	174
	Above Average	10	23	37	70
		119	163	118	400

$H_0 : \theta_{ij} = \theta_i \theta_j$, that is, the probability that someone falls in the ij cell is the product of the two probabilities, i.e., job performance and IQ are independent

Probability & Statistics for Computer Science

- e_{ij} is the expected number in cell (ij)
- f_{ij} is the observed number in cell (ij)
- f_i is the row total and f_j is the column total
- f is the grand total

Probability & Statistics for Computer Science

Finally we have

$$e_{ij} = \frac{f_i \cdot f_j}{f}$$

and the test statistic

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(f_{ij} - e_{ij})^2}{e_{ij}}$$

We reject H_0 if $\chi^2 > \chi^2_{\alpha,(r-1)(c-1)}$

Probability & Statistics for Computer Science

Lets’ solve this...

		Performance			
		Poor	*Fair*	*Good*	
	Below Average	67	64	25	156
I.Q.	*Average*	42	76	56	174
	Above Average	10	23	37	70
		119	163	118	400

References

Miller, I. and Miller, M. (2014).
John E. Freund's Mathematical Statistics with Applications.
Pearson, 8th edition.