Lecture 4: Two-sample tests

BMI 713
October 31, 2017
Peter J Park

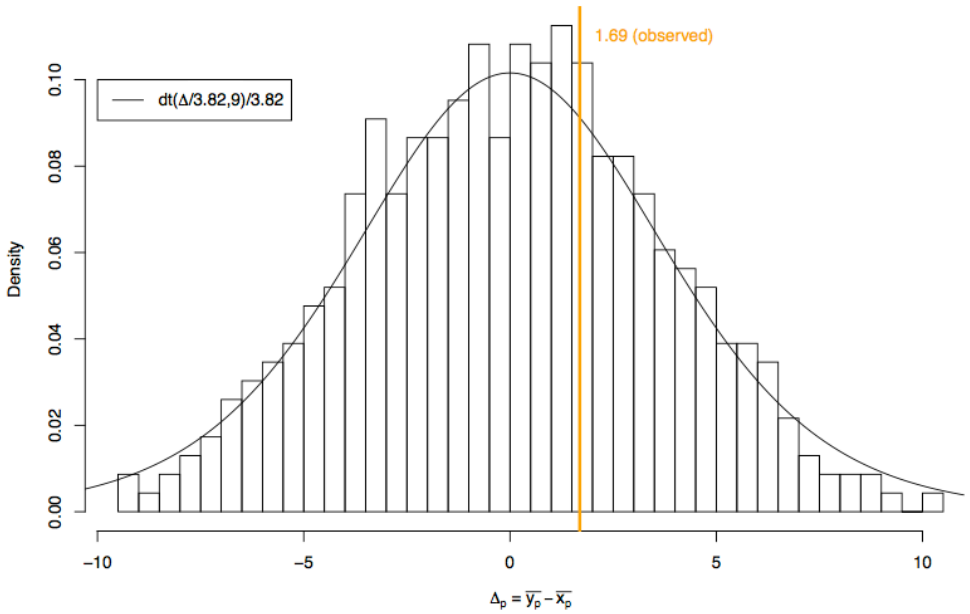
"Student's t-test"

Last time:

- t-test for one-sample testing

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \qquad t = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

- Permutation testing



Two-sample test

- Previously, we compared a sample mean with a known value:

$$H_0 : \mu = \mu_0$$

- Now, we would like to compare two unknown means:

$$H_0 : \mu_1 = \mu_2$$

- Are the data **paired or unpaired**?
- Two samples are paired if each data point in one sample is related to a unique data point in the other.
- Example: testing each patient before and after intervention

Paired data

- For paired data, this is really just a one-sample test of the differences $d_i = x_{i2} - x_{i1}$ with mean Δ.

$$H_0 : \mu_1 = \mu_2 \quad \longrightarrow \quad H_0 : \Delta = 0$$

- So the test statistic

$$t = \frac{\bar{d}}{s_d / \sqrt{n}}$$

where $\bar{d}$ is the sample average of the differences and s_d is the sample s.d. of the differences, follows a *t*-distribution with *n-1* degrees of freedom (d.o.f.) under H_0

Unpaired two-sample test

- Two versions, depending on whether the variances in the two populations are equal or not.

Unequal Variances

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

t follows a *t*-distribution with some d.o.f. (the formula is complicated)

Equal Variances

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

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

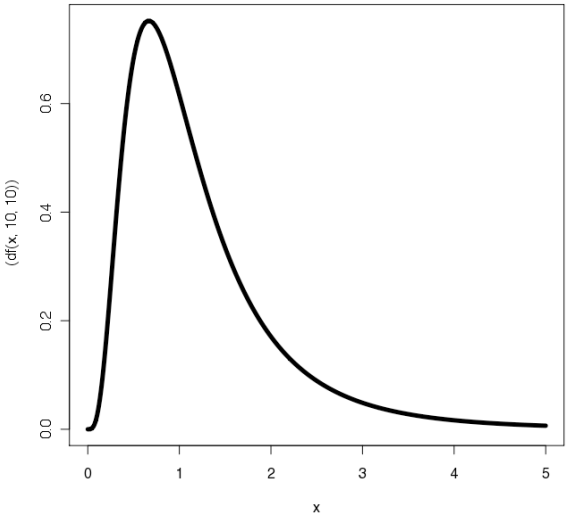
t follows a *t*-distribution with (n_1+n_2-2) d.o.f.

Unpaired two-sample test

- How to decide whether the variances are the same or not?
- Using unequal variances should be fine in most cases.
- F-test to test if the populations have equal variances
- Consider two normal populations with means μ_1 and μ_2 and a common variance σ^2. If you take two samples of size n_1 and n_2,

$$F = \frac{s_1^2}{s_2^2}$$

follows an F-distribution with (n_1-1, n_2-1) d.o.f.



Example

- Is daily caloric consumption equal in two populations?
- We took two samples and found the following:

$$\bar{X}_1 = 2500, s_1 = 250, n = 13$$
$$\bar{X}_2 = 2000, s_2 = 200, n = 10$$

- Test the hypothesis that the mean consumption is the same in both populations.

Example

- Step 1: Test equality of variances.

$$F = \frac{s_{max}^2}{s_{min}^2} = \frac{250^2}{200^2} = 1.56$$

$$F_{0.975}(13-1,10-1) = 3.87$$

qf(.975,12,9)

- There is no evidence for different population variances ➔ use equal variances method.

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

$$= \frac{(13-1)250^2 + (10-1)200^2}{13+10-2} = 52857 = 229.9^2$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{500}{229.9\sqrt{\frac{1}{13}+\frac{1}{10}}}$$

$$= 5.17 > 2.080 = t_{21,0.975}$$

- We reject the null hypothesis and conclude that the two populations have different means.

Confidence Interval

- The 100(1-α)% confidence interval for the true mean differences?
- Recall that we expected 95% of the data from a normal distribution to be contained in $\mu \pm 1.96\sigma$.

Unequal Variances

$$(\bar{x}_1 - \bar{x}_2) \pm t_{d',\,1-\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Equal Variances

$$(\bar{x}_1 - \bar{x}_2) \pm t_{n_1+n_2-2,\,1-\frac{\alpha}{2}} \left(s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right)$$

- So in the current example, the CI is

$$(2500 - 2000) \pm t_{21,0.975} \left(229.9 \sqrt{\frac{1}{13} + \frac{1}{10}} \right)$$

$$500 \pm 2.080(96.7) = 500 \pm 201$$

Other Methods for Two-group Comparisons

- Many methods have been devised for this problem.

- Where can the t-test go wrong?
  - With a small sample, *s* may be under-estimated by chance
  - "Regularization"

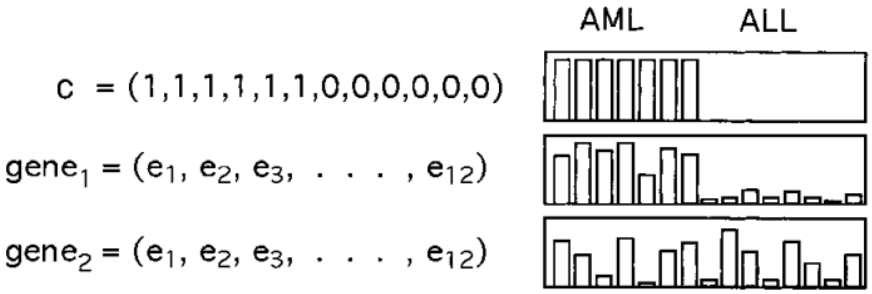
$$t = \frac{\bar{X} - \mu}{s_0 + s/\sqrt{n}}$$

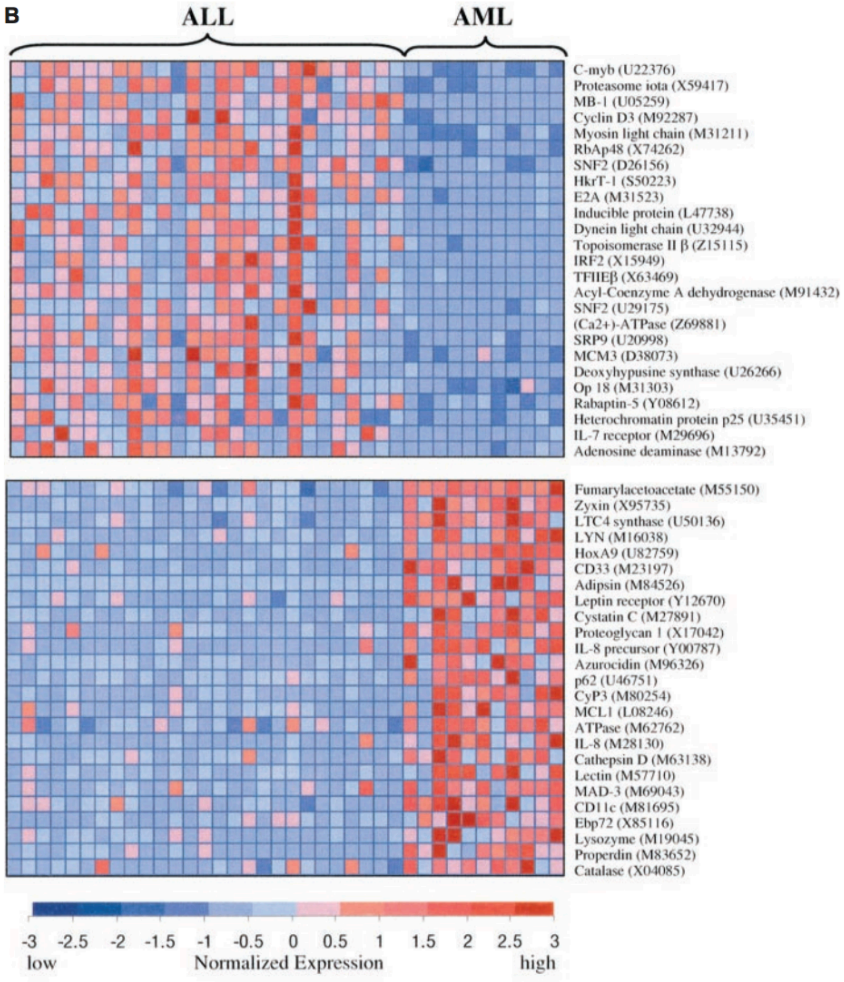
- Another option is to use a Bayesian method.

Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring

T. R. Golub,[1,2*†] D. K. Slonim,[1†] P. Tamayo,[1] C. Huard,[1] M. Gaasenbeek,[1] J. P. Mesirov,[1] H. Coller,[1] M. L. Loh,[2] J. R. Downing,[3] M. A. Caligiuri,[4] C. D. Bloomfield,[4] E. S. Lander[1,5*]

Science, 1999, cited 12000+ times





"signal to noise ratio"

$$P(g,c) = [\mu_1(g) - \mu_2(g)]/[\sigma_1(g) + \sigma_2(g)]$$

ANOVA

- What if we have more than two groups?
- Suppose we have *k* populations, each roughly normal with common variance σ^2.
- How do we test for H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$?
- The extension of the t-test to this case is known as one-way **Analysis of Variance**.
- The name is deceptive: we need to analyze variances to test for a difference in means.
- What is $\mathbf{H_1}$? **That at least one of the population means differs from one of the others.**

Inference on proportions

- Recall: Binomial Distribution
- Binomial distribution
  - Two categories: “success” and “failure”
  - Each trial is independent with probability *p*
  - A fixed number of trial
- If a random experiment has two possible outcomes and we do *n* independent repetitions with identical success probability *p*, then $X \sim \text{Bin}(n,p)$ and

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

Binomial Distribution

- "Exact" methods calculate sum of the discrete probabilities to compute p-values.
- This is not feasible or necessary for large sample sizes.
- In such cases, we use the normal approximation to the binomial distribution.

- What are the mean and variance of the normal approximation?

$$E(X) = np$$

$$Var(X) = np(1-p)$$

Estimation of a Population Proportion

- Example: We would like to estimate p, the probability that a person under the age of 40 who is diagnosed with lung cancer survives for at least 5 years.
- A random sample of $n = 70$ individuals is selected from the population.
- It is found that only $X = 8$ patients out of the 70 survive for 5 years.
- The number of "successes" X has a **binomial distribution**.
- Hypothesis testing can be done using the "exact" method for small sample sizes.

Sampling Distribution of Proportions

- The 5-year survival probability is estimated by the sample proportion of individuals who survive for 5 years

$$\hat{p} = \frac{X}{n} = \frac{8}{70} = 0.114$$

- If repeated samples of size 70 are selected from the population, what is the **sampling distribution of proportions**?
- The mean: $$E(\hat{p}) = p$$
- The variance: $$Var(\hat{p}) = \frac{pq}{n}$$

Normal Approximation

- We apply the central limit theorem to the binomial distribution.
- If we draw samples of size *n* from a population whose proportion of interest is *p*, then the sample proportions *p* will be approximately distributed as

$$\hat{p} \sim N(p, pq/n)$$

provided that *n* is large enough, e.g., $npq \geq 5$

Since *p* and *q* are not known, replace them with $\hat{p}$ and $\hat{q}$

Back to the Example

- For the lung cancer 5-year survival data

$$n\hat{p}\hat{q} = 70(0.114)(0.886) = 7.1$$

- Since this is large enough, the distribution of $\hat{p}$ can be assumed to be normal:

$$z = \frac{\hat{p} - p}{\sqrt{pq/n}}$$

  is distributed approximately as N(0,1)

Confidence Interval

- 95% confidence interval:

$$P(-1.96 \le z \le 1.96) = 0.95$$

- We substitute *z*

$$P\left(-1.96 \le \frac{\hat{p} - p}{\sqrt{pq/n}} \le 1.96\right) = 0.95$$

- Isolating *p* in the center

$$P\left(\hat{p} - 1.96\sqrt{\frac{pq}{n}} \le p \le \hat{p} + 1.96\sqrt{\frac{pq}{n}}\right) = 0.95$$

- Therefore, the 95% confidence interval is

$$\left(\hat{p} - 1.96\sqrt{\frac{pq}{n}}, \hat{p} + 1.96\sqrt{\frac{pq}{n}}\right)$$

Confidence Interval

- Since p is not known, we estimate this with

$$\left(\hat{p}-1.96\sqrt{\frac{\hat{p}\hat{q}}{n}},\hat{p}+1.96\sqrt{\frac{\hat{p}\hat{q}}{n}}\right)$$

- For the proportion of individuals under the age of 40 who survive at least 5 years after being diagnosed with lung cancer is

$$\left(.114-1.96\sqrt{\frac{(.114)(.886)}{70}},.114+1.96\sqrt{\frac{(.114)(.886)}{70}}\right)$$

$$(0.041,0.188)$$

Hypothesis Testing for One Proportion

- To test the null hypothesis

$$H_0 : p = p_0,$$

if $np_0q_0 \geq 5$ $(q_0 = 1 - p_0)$ then under H_0 the test statistic

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}}$$

is approximately normally distributed as N(0,1)

Example

- Mendel: self-pollination of a pea plant that was heterozygous (Dd) for the dwarf gene would yield 3/4 tall plants and 1/4 dwarf plants. Among the F2 progeny from a cross of a tall (DD) and a dwarf (dd) plant, Mendel observed 787 tall plants and 277 dwarfs.

H0: p=0.75

$$Z = \frac{787/1064 - .75}{\sqrt{\frac{.75(1-.75)}{1064}}} = -0.7788$$

Thus, the p-value is 0.436 and we do not reject H_0

Hypothesis Testing for Two Proportions

- Suppose two populations have unknown proportions p_1 and p_2 and we want to test

$$H_0 : p_1 = p_2$$
$$H_A : p_1 \neq p_2$$

- Take two samples of size n_1 and n_2, compute. $\hat{p}_1$ and $\hat{p}_2$
- If H_0 is true, then both populations have the same proportion p, which we estimate as

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

Proportion Test

- If $n_1\hat{p}_1(1-\hat{p}_1) \geq 5$ and $n_2\hat{p}_2(1-\hat{p}_2) \geq 5$

  the under H_0, the test statistic

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

  is distributed approximately as N(0,1)

Example

- Pagano & Gauvreau Ch14
- An 18-month study on effectiveness of seat belts on mortality among pediatric victims of motor vehicle accidents.
- H_0: the proportions of children who die as a result of the accident are identical for the two groups wearing and not wearing seat belts.
- In the sample of 123 wearing a seat belt, 3 died.
- In the sample of 290 not wearing a seat belt 13 died.

Example

$$\hat{p}_1 = x_1 / n_1 = 3/123 = 0.024$$
$$\hat{p}_2 = x_2 / n_2 = 13/290 = 0.045$$

- Is the discrepancy in sample proportions too large to be attributed to chance?

$$\hat{p} = (3+13)/(123+290) = 0.039$$

$$z = \frac{(0.024 - 0.045) - 0}{\sqrt{(0.039)(1-0.039)\left(\frac{1}{123} + \frac{1}{290}\right)}} = -1.01$$

- The p-value is 0.312. Therefore, we cannot reject the null hypothesis. This study does not provide evidence that the proportions of children dying differ between those who were wearing set belts and those who were not.