Multivariate Stats, Week 10

Noah Silbert

October 27, 2017

To do

1. homework review
2. comparing (multidimensional) means

Recap: Maximum likelihood estimates

If we have a set of independent, identically distributed (iid) normal random variables $\mathbf{x} = \{x_1, x_2, ..., x_n\}$, and we let the parameters μ and σ vary, then we have the likelihood function:

$$f(\mathbf{x}|\mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)$$
$$= \frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^n} \exp\left(-\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

The values of μ and σ that maximize this function are called *maximum likelihood estimates*. If $\mathbf{x}$ are normally distributed, then $\bar{x}$ (the sample mean) is the MLE for μ, and $s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \hat{\mu})^2}{n}}$ is the MLE for σ. **Click here for more information**.

Recap: the multivariate normal likelihood

As we've seen, the multivariate normal density function for data vector $\mathbf{x}$ of length p (i.e., with p dimensions) is:

$$f(\mathbf{x}|\boldsymbol{\mu}, \Sigma) = \frac{1}{\sqrt{2\pi}^p |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})\right)$$

The likelihood for n observations (i.e., a $n \times p$ data matrix $\mathbf{X}$) is:

$$\begin{aligned} f(\mathbf{X}|\boldsymbol{\mu}, \Sigma) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}^p |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu})\right) \\ &= \frac{1}{\sqrt{2\pi}^{np} |\Sigma|^{n/2}} \exp\left(-\sum_{i=1}^{n} \frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu})\right) \end{aligned}$$

Recap: multivariate maximum likelihood estimation

- $\bar{\mathbf{x}}$ is the maximum likelihood estimate of $\boldsymbol{\mu}$, and $\mathbf{S} = \frac{1}{n}\mathbf{x}^T\left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right)\mathbf{x}$ is the maximum likelihood estimate of Σ.
- If $\mathbf{x}$ is $N_p(\mu, \Sigma)$, then $\bar{\mathbf{x}}$ is $N_p(\mu, \Sigma/n)$. If $\mathbf{x}$ is non-normal, then $\bar{\mathbf{x}}$ is approximately $N_p(\mu, \Sigma/n)$, as long as n is large enough.
- $\mathbf{W} = \mathbf{x}^T\left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right)\mathbf{x}$ is a Wishart random variable (i.e., a multivariate generalization of χ^2) with $n-1$ degrees of freedom and scale matrix $\mathbf{V}$.
- $\bar{\mathbf{x}}$ and $\mathbf{S}$ are statistically independent.

One-sample test of a univariate mean

A one-sample t statistic provides information about the statistical distance between an observed sample mean and a hypothesized true mean.

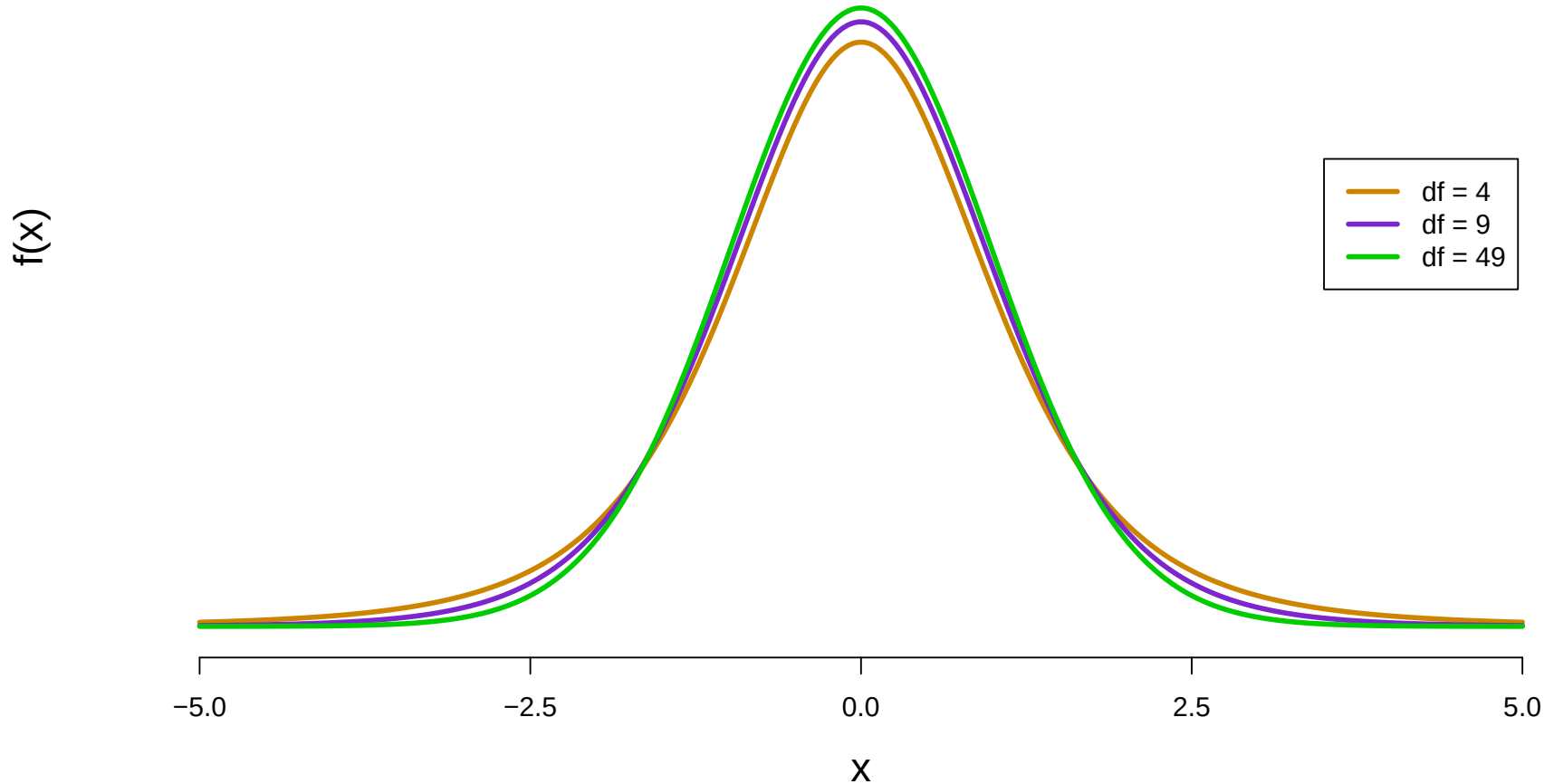
The t statistic can be used to test the null hypothesis $H_0 : \mu = \mu_0$.

More specifically, the t statistic is: $t = \dfrac{\sqrt{n}(\bar{x} - \mu_0)}{s} = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$.

Note that the denominator shrinks as n increases - a larger sample makes estimation more precise.

t distributions

The difference between two normally-distributed means would be normally distributed, but we divide by an estimate of the standard deviation, which gives us the t distribution:



One-sample multivariate generalization of the t test

The multivariate generalization of the t test looks very similar. In this case, the null hypothesis is $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$.

The test statistic for the one-sample multivariate comparison of means is:

$$\begin{aligned} T^2 &= n\left(\bar{\mathbf{x}} - \boldsymbol{\mu}_0\right)^T \mathbf{S}^{-1} \left(\bar{\mathbf{x}} - \boldsymbol{\mu}_0\right) \\ &= \left(\bar{\mathbf{x}} - \boldsymbol{\mu}_0\right)^T \left(\frac{\mathbf{S}}{n}\right)^{-1} \left(\bar{\mathbf{x}} - \boldsymbol{\mu}_0\right) \end{aligned}$$

Note that this is a scalar, squared statistic, so it only takes positive values (and is right skewed). The statistician Harold Hotelling worked out the distribution function for this statistic, which allows us to calculate p values for observed T^2 statistics.

Two-sample univariate comparison of means

A two-sample t statistic provides information about the statistical distance between the means for two groups.

The t statistic can be used to test the null hypothesis $H_0 : \mu_1 = \mu_2$.

More specifically, the t statistic is: $t = \dfrac{\bar{x}_1 - \bar{x}_2}{s_{pl}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$.

Here, s_{pl} is the pooled within-group standard deviation.

In R, you can do one- or two-sample t-tests with the function `t.test()`.

Multivariate comparison of two mean vectors

The multivariate test of $H_0 : \mu_1 = \mu_2$ uses the following test statistic:

$$T^2 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T \left[\left(\frac{1}{n_1} + \frac{1}{n_2} \right) S_{pl} \right]^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$$

As we have seen before, $\mathbf{S}_{pl} = \dfrac{1}{n_1 + n_2 - 2} (\mathbf{W}_1 + \mathbf{W}_2)$, where $\mathbf{W}_i = (n_i - 1)\mathbf{S}_i$

Conveniently, there is an R package called 'Hotelling' that makes it easy to carry out a two-sample multivariate "t test." (See associated R script.)

Comparing paired mean vectors

If the two "groups" you are interested in testing are paired (e.g., one group is pre-test, the other is post-test), you can simply take the *difference* of each observation vector and do a multivariate test of $H_0 : \boldsymbol{\mu_d} = 0$.

For example, suppose $\mathbf{X}_a$ is the data matrix for the pre-test and $\mathbf{X}_b$ is the data matrix for the post-test. Let $\mathbf{D} = \mathbf{X}_a - \mathbf{X}_b$, and calculate $\bar{\mathbf{d}}$ (i.e.,`colMeans(D)`) and $\mathbf{S_D}$ (`cov(D)`).

Then $T^2 = n\bar{\mathbf{d}}^T \mathbf{S_D}^{-1} \bar{\mathbf{d}}$.

LDA, logistic regression, and Hotelling tests

Note that with LDA and logistic regression, we were using a set of continuous variables to predict the value of a discrete variable (e.g., a group label).

With Hotelling's test to compare two multivariate means, we are essentially just swapping the roles of the set of continuous variables and the discrete variable, using the discrete variable to predict (the difference between the mean vectors of) the continuous variables.

Similarly, when we use MANOVA, it will be similar to LDA with more than 2 categories, just with the independent and dependent variables switched.