Advanced Statistics for Neuroscience Exam preparation

Topics

1. Mean, median, mode (10pt)
   - Be able to calculate by hand
2. Distributions (10pt)
   - Names, values integers or not, positive or not, infinite or not, one or two parameters
3. Normally distributed samples (10pt)
   - Formulas for mean, variance, etc.
4. T-test, F-test, Chi-square-test, KS-test (10pt)
   - What does each test, what properties do they compare
5. Measures for pairwise association (10pt)
   - formulas for Pearson, Rank-Order, Chi-square, mutual information
6. Linear regression (10pt)
   - Formulas for sum of squared errors and estimated variance, interpretation of log likelihood and BIC
7. Logistic regression (10pt)
   - probabilities and logits, construction of design matrices, formulat for binomial variance
8. FDR correction (10pt)
   - Apply Benjamini-Hochberg criterion to p-values
9. Boostratpping (10pt)
   - Explain computational steps of bootstrapping
10. Linear discriminant analysis (10pt)
    - explain computational steps of LDA and potentially formulas
11. Formulas

Mean, median, mode

Mean is the average of a set. *Median* is the point that divides a distribution into two areas of equal size, i.e. the middle point (if dataset has an even number of points, take the average of the two center values). *Mode* is the most probable point (the most frequent number in the dataset).

Variability, how spread out are the observations in a distribution. Typically measured as a distance from the mean. *Variance* is the average squared difference from the mean.

Mean:

$$\mu_x = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Variance, true variance can only be calculated when the true mean (μ) of the distribution is known:

$$Var(x) = \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \langle x \rangle)^2$$

Standard deviation is the square root of variance:

$$Std(x) = \sqrt{Var(x)}$$

Distributions

Unimodal, bimodal, multimodal **be careful not to confuse *bimodal* and *binomial***

Normal distribution - a Gaussian distribution with zero-mean and unit variance, symbol for Normal distribution is Φ. Can include positive or negative numbers. Dependent on the parameters μ (mean) and σ (standard deviation, σ^2 is variance)

Gamma distribution - a two parameter distribution that includes the shape parameter α and scale parameter θ (or instead of the scale parameter, the rate parameter $\lambda = 1/\theta$). Exponential and chi-square distributions are examples of gamma distributions. Only include positive numbers.

Lognormal distribution - grows in multiplicative scale, always takes positive values. Many biological observations happen in lognormal distribution

Bernoulli distribution- a distribution describing the outcome of a binary experiment. A special case of the binomial distribution.

Poisson distribution - "the law of rare events", the probability of a given number of events happening in a fixed interval of time. Defined only at positive integer values, where λ is the expected rate of occurence. As λ grows, the curve gets lower and more spread out.

When the area to the left of the mode is larger, it is called 'negative skew', when the right area to the right of mode is larger, it is 'positive skew'

Chi-square distribution χ^2

Normally distributed samples

Mean and variance of normally distributed data are listed in *Section 1.*

Additionally, can incorporate Bessel's correction, a means for correcting for sample variance where we have n observations x_i but no further information (whereas in the original variance we have the *true population mean* μ), this incorporates $\frac{1}{N-1}$ such that we get a corrected sample variance of:

$$m = \frac{1}{N}\sum_i x_i, \qquad s^2 = \frac{1}{n-1}\sum_i (x_i - m)^2$$

For normally distributed samples, the expected value of Bessel's correct variance equals true variance.

Standard error calculates the standard deviation of the sample mean. It's value is inferred, not observed, since it is extrapolating from a hypothetical scenario where we repeatedly collect new samples many times, it is calculated for true variance as:

$$SE(x) = \frac{Std(x)}{\sqrt{n}} = \frac{\sigma}{\sqrt{n}}$$

But when variance and mean are calculated from sample population, $SE(x)$ is calculated as:

$$SE(x) = \sqrt{\frac{1}{n(n-1)}\sum_i (x_i - m)^2}$$

Z-Scoring is when we normalize a dataset to the mean and standard deviation, and is useful for identifying how far a sample deviates from the mean. Calculated as such:

$$z_i = \frac{x_i - \mu}{\sigma}$$

T-test, F-test, Chi-square-test, KS-test

T-test is a test for whether two populations differ under the null distribution.

$$t_m = \frac{m - \mu_0}{s^2/\sqrt{n}}, \qquad m = \frac{1}{n}\sum_i x_i, \qquad s^2 = \frac{1}{n-1}\sum_i (x_i - m)^2$$

Where μ_0 is the hypothetical true mean that we are testing our sample mean against.

The Student's t-distribution tells us how an observed sample mean is distributed around the true mean when we don't know the true variance. We can use the t-test for performing hypothesis testing, i.e. that the hypothesis is either: *due to chance* (the null hypothesis H_0) or *not due to chance* (alternate hypothesis H_1). This requires us to select a level of significance, known as the α, often something like $\alpha = 0.05$ (5%) or lower. In a normal distribution, 95.4 of the probability

mass function falls within $\pm 2\sigma$ of μ. Typically the p-value is selected from a table that correlates p-values to t-statistics

Independent (unpaired) t-test is used when we have two independent sample sets–such as wild-type and mutants–and want to test whether the means are significantly different. Calculate the sample means and variance for each group as above, then calculate the standard error of the difference as:

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Where s_1^2 and n_1 are the sample variance and n for sample 1, and s_2^2 and n_2 are the sample variance and n for sample 2. The test statistic for the difference is then:

$$t_{diff} = \frac{(m_1 - m_2) - (\mu_1 - \mu_2)}{SE}$$

Which, for the null hypothesis that $\mu_1 = \mu_2$ becomes:

$$t_{diff} = \frac{(m_1 - m_2)}{SE}$$

With $n_1 + n_2 - 2$ degrees of freedom

Paired (dependent) t-test is used when we have paired samples, such as observations of the same n individuals before and after treatment. Must first calculate the average distance between sample pairs:

$$m_D = \frac{1}{n} \sum_{i=1}^{n} d_i, \qquad d_i = x_{1i} - x_{2i}$$

Where x_{1i} and x_{2i} are the same subject between two conditions. Then the estimated variance becomes:

$$s_D^2 = \frac{1}{n-1} \sum_{i=1}^{n} (d_i - m_D)^2$$

and the standard error of differences becomes:

$$SE_D = \frac{s_D}{\sqrt{n}}$$

Giving us a test statistic of:

$$t_{diff} = \frac{m_D}{SE_D} = \frac{m_D}{s_D/\sqrt{n}}$$

F-test measures for differences in variance. Two distributions may have similar means but different variances. The F-statistic is the ratio of the two sample variances:

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

Chi-square test allows us to infer whether two sample sets are drawn from the same or different distributions. Used for **binned** datasets, i.e. **will only accept positive integers**. For continuous data, we use the *Kolmogorov-Smirnov test (KS)*.

$$\chi^2 = \sum_{i=1}^{N} \frac{(n_i - e_i)^2}{e_i}$$

Where N is the number of distinct categories the observations are binned into, i is all of the integers between $\{1, ..., N\}$, n_i is the number observed in each bin, and n_e is the number expected per bin.

Measures for pairwise association

Parametric tests assume normal distribution, non-parametric tests don't.

Pearson's r - also known as linear correlation coefficient. Given paired observations (x_i, y_i), with $i = 1, 2, \ldots, N$, r is calculated as:

$$r = \frac{(x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_i (x_i - \mu_x)^2}\sqrt{\sum_i (y_i - \mu_y)^2}}$$

Linear regression

Logistic regression

FDR correction

Boostrapping

Linear discriminant analysis

Formulas

Normal Distribution

$$\Phi(u) = \frac{1}{\sqrt{2\pi\sigma^2}} exp^{\frac{(x-\mu)^2}{2\sigma^2}}$$