# Introduction to Statistical Data Science

November 26, 2018

## 1 Lecture 1

### 1.1 Normal Distribution

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

$$F(x) \equiv P(X \leq x)$$

**Central Limit Theorem (for normal distribution)**

The more random variables we average over, the closer the resulting distribution will be to the Normal distribution

**Parameter:** $u$ The mean is the location parameter.

**Parameter:** $\sigma^2$ The variance is the scale parameter

### 1.2 Uniform Distribution

$$X \sim U[0,1] \tag{1}$$

$$0 \leq x \leq 1 \tag{2}$$

$$p(x) = 1 \tag{3}$$

$$F(x) = P(X \leq X) = x \tag{4}$$

$$\tag{5}$$

**Use uniform distribution to construct normal distribution**

$$\mathbf{X} = \begin{bmatrix} X^{(1)} \\ X^{(2)} \\ \cdots \\ X^{(n)} \end{bmatrix} \tag{6}$$

$$X^{(i)} \sim U[0,1]$$

$$Y = \frac{1}{n}\sum_{i=1}^{n} X^{(i)}$$

subsampling $\mathbf{X}$ vector to construct $Y$. The distribution of $Y_j \sim ?, j = 1, 2, 3, \ldots, p$ would close to normal distribution.

### 1.3 Poisson distribution

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

$$E(X) = V(X) = \lambda$$

### 1.4 empirical CDF

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(x^{(i)} \leq x)$$

quantile is simply the inverse of the CDF: -The 0.9 quantile is the value of $x$ such that $F(x) = 0.9$ i.e. $x = F^{-1}(0.9)$

## 2 Hypothesis

### null hypothesis

In inferential statistics, the null hypothesis is a general statement or default position that there is no relationship between two measured phenomena, or no association among groups.

### p-value

The probability of obtaining results as or more extreme than that observed, assuming $H_0$ is true, is the p-value, under the assumption that the null hypothesis.

$$p \equiv P(X \leq 15; H_0)$$

$$p = \sum_{x=0}^{15} \binom{40}{x} 0.5^x (1-0.5)^{(40-x)}$$

The p-value is most certainly not the probability of $H_0$ being true

When used in practice with a threshold of 0.05 this is an informal method of reasoning and can be easily criticized

## Statistical power

The power of a hypothesis test is the probability of avoiding a false negative

## P-value distribution (CDF)

$$P(F(X) \leq z) = P(F^{-1}(F(x)) \leq F^{-1}(z)) = P(X \leq F^{-1}(z)) = F(F^{-1}(z)) = z$$

### Level

The threshold probability of 0.05 was the level of the test.

The choice of a particular level may be guided by the need to trade off Type 1 and Type 2 errors.

### Type 1 errors

Type 1 error occurs when we reject the null hypothesis $\mathrm{H}_0$, when it is true

### Type 2 errors

Type 2 error occurs when we fail to reject $\mathrm{H}_0$ when it is false.

### power of the test

The probability of avoiding a Type 2 error is the power of the test.

That is the probability that we reject $\mathrm{H}_0$ given that it is false.

The power of a test varies with sample size

The power of a test also varies with the level of the test.

The ways to increase the power of our test: collect more data, allow for a higher Type 1 error, use a better test statistic, make stronger assumption.

### Testing procedure

Specify a null and alternative hypothesis.

Specify the level of the test.

Specify a suitable test statistic.

### critical region

The set of all test statistic values which would cause us to reject $\mathrm{H}_0$.

### t-test

For the small sample, and iid normal distribution:

$$T = \frac{\sqrt{n}(\bar{X}_n - \mu_0)}{S_n} \sim \tau(n-1)$$

This is called a t-distribution with $n-1$ degrees of freedom

### Wald test

Sample $X^{(i)}$ can be from any distribution.

Sample sizes must be assumed to be 'big enough' that CLT applies.

Hence, the distribution of test statistic is $\mathrm{N}(0,1)$. That is the case of t-distribution which $n \to \infty$

### Goodness-of-fit tests (chi-square test)

still needs to be discussed.

### Paired tests

## Bonferroni correction

While testing $n$ independent hypothesis in the same data set, the p value should be $1/n$

## Confidence Intervals

$$X \sim N(\mu, \sigma^2)$$

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$\bar{X}$ is the sample average. If $X_i \sim N(\mu, \frac{\sigma^2}{n})$, then

$$\bar{X} \sim N$$

$$Var(\bar{X}) = \frac{(\frac{\sum_i x_i}{n} - \mu)^2}{1} = (\frac{\sum_i (x_i - \mu)}{n})^2 = \sigma^2/n$$

Then

$$\frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \sim N(0, 1)$$

This a pivot, a function of parameter of interest which has a known distribution.

$$P\left(\frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \leq 1.5\right) = p(Z \leq 1.5) = p\left(\mu \geq \bar{X} - \frac{1.5\sigma}{\sqrt{n}}\right) = 0.93$$

**Coverage**

The interval $[\bar{X} - 1.5\sigma/\sqrt{n}, \infty)$ will contain $\mu$ 93% of the time

## two side interval

It is hard to calculate the confidence interval, therefore a CLT could help t solve it with large sample.

## Bootstrap

A foundation of bootstrap is sampling with replacement

1. Randomly select a data point.

2. Add it to the 're-sample'

3. Put it back in the box.

## Using Bootstrap to calculate SE

1. Draw $X^{(1)}*, \ldots, X^{(n)}* \sim \hat{F}_n$

2. Compute $\hat{X}_n^*$ by averaging $X_1^*, \ldots, X_n^*$

3. Repeat steps 1 and 2, $B$ times, to get $\bar{X}_{n,1}^*, \ldots, \bar{X}_{n,B}^*$

4. Let

$$s.e._{boot} = \sqrt{\frac{1}{B}\sum_{b=1}^{B}\left(\bar{X}_{n,b}^* - \frac{1}{B}\sum_{r=1}^{B}\bar{X}_{n,r}^*\right)^2}$$

### Bootstrap pivotal interval

Define $H(r) = P(\hat{\theta} - \theta \leq r)$

Define quantiles such that we get coverage $1 - \alpha$:

$$\begin{aligned} P(a(\hat{\theta}_n) \leq \theta \leq b(\hat{\theta}_n)) &= 1 - \alpha \\ a(\hat{\theta}_n) &= \hat{\theta} - H^{-1}(1 - \alpha/2) \\ b(\hat{\theta}_n) &= \hat{\theta} - H^{-1}(\alpha/2) \end{aligned}$$

## Linear Regression

Data points satisfies: $y_i = \beta_0 + \beta_1 x_i + e_i (i = 1, \ldots, n)$

Next, we are trying to find $\bar{y} - \hat{\beta}_0 - \hat{\beta}_1 \bar{x} = 0$ and $\sum_{i=1}^{n} x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$ The sum of squared prediction errors:$RSS = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$

Then take the derivative of parameter $\hat{\beta}_0$ and $\hat{\beta}_1$ as 0, and get the result.

After substituting the result, it could be concluded that

$$\begin{aligned} \hat{\beta}_1 &= [\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})]/[\sum_{i=1}^{n}(x_i - \bar{x})^2] \\ \hat{\beta}_1 &= \frac{C_{xy}}{C_{xx}} \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ &\text{where} \\ C_{xx} &= \sum_{i=1}^{n}(x_i - \bar{x})^2 = (\sum_{i=1}^{n} x_i^2) - n\bar{x}^2 \\ \text{and} \quad C_{xy} &= \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = (\sum_{i=1}^{n} x_i y_i) - n\bar{x}\bar{y} \\ RSS &= C_{yy} - \frac{C_{xy}^2}{C_{xx}} \end{aligned}$$

### Error

If $x_i = x_j$, usually $y_i \neq y_j$. Therefore treat $y_i$ as realized value of random variable $Y_i$, whose distribution depends on $x_i$.

The model would be: $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ where $\epsilon_i$ are independent random variables with $E(\epsilon) = 0, Var(\epsilon) = \sigma^2$

# lecture 4

## Dispersion

Dispersion is a broad concept relating to the variability in a distribution.

## Logistic Regression

### Linear Regression

$$\begin{aligned}\mathbb{E}[Y_i|\eta_i] &= \mu_i = \eta_i \\ Y_i|\eta_i &\sim N(\eta_i, \sigma^2) \\ Y_i &\in \mathbb{R}\end{aligned}$$

### Logistic Regression

$$\begin{aligned}\mathbb{E}[Y_i|\eta_i] &= \mu_i = \theta_i = g^{-1}(\eta_i) \\ Y_i|\eta_i &\sim Bernoulli(g^{-1}(\eta_i)) \\ \eta_i &\in \mathbb{R}, \theta_i \in [0,1]\end{aligned}$$

The link function $g^{-1}(\eta_i)$ would be a logistic function

$$\mu_i = g^{-1}(\eta_i) = logistic(\eta_i) = \frac{1}{1+\exp(-\eta_i)}$$

$$\eta_i = g(\mu_i) = logit(\mu_i) = \log(\frac{\mu_i}{1-\mu_i}) = \log(\frac{\mathbb{P}(Y_i = 1|\eta_i)}{\mathbb{P}(Y_i = 0|\eta_i)})$$

Possible link function would be $g(\mu_i) = \Phi^{-1}(\mu_i)$, the inverse of the standard normal CDF, and the complimentary log-log function, $g(\mu_i) = \log(-\log(1-\mu_i))$.

A similar concept for generalized linear models is called the deviance ($R^2$ was a measure of model fit for normal linear regression models)

$$D(M_1, M_0) = 2\log(\frac{L(\hat{\beta}^{(1)}; x, y)}{L(\hat{\beta}^{(0)}; x, y)}) = 2(I(\hat{\beta}^{(1)}; x, y) - (\hat{\beta}^{(0)}; x, y))$$

Asymptotically, for $M_{p2}$ with $p2$ predictors nested within model $M_{P1}$ with $p_1 > p_2$ predictors:

$$D(M_{p1}, M_{p2}) \sim \chi^2_{p1-p2}$$

### Poisson regression

$$\begin{aligned} \mathbb{E}[Y_i|\eta_i] = \mu_i = \eta_i \\ Y_i|\eta_i \sim Poisson(g^{-1}(\eta_i)) \\ \eta_i \in \mathbb{R}, \theta_i \in [0, \infty] \end{aligned}$$

## Combinatorial search (subset selection)

search among possible subsets of inputs, fit final model based on the chosen subset.