22/2/2022

MATH 104 L01

Chapte 1 : Introduction to Data

all variables

(chp 2+3) numerical — continous, discrete

categorical (chp 4) — regular categorical, ordinal

Association vs Causation

FP2
Chp 3

MATH 104 LO2

Normal Distribution (2.3) -

The normal distribution model always describes a symmetric, unimodal, bell-shaped curve. changing the mean moves the graph left $\leftrightarrow$ right and changing standard deviation stretches or constricts the curve.

If a Normal Distribution has mean μ and sd σ we can write it $N(\mu, \sigma^2)$ The mean and sd are called the distribution's parameters.

PDF = $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$

CDF = $P(X \leq x)$ $F(x) = \int_{-\infty}^{x} f(u)\,du$ Doesn't have an analytical solution

denoted pnorm(x, μ, σ)

$N(0, 1^2)$ is the standard normal distribution special case where CDF $F(x) = \phi(x)$ denoted pnorm(x)

Z-score $Z = \frac{x-\mu}{\sigma}$ used to convert a normal to standard normal and then compute on calculator /R.

pnorm(x) $\rightarrow$ quantile / percentile as it calculates an area under the distribution.

R example -

1) pnorm(x, μ, σ) * n sub in values. n=100
2) $Z = (x-\mu)/\sigma$
pnorm(Z)

option 1) or 2)

pnorm(x) → quantile (q)

$Z = \text{qnorm}(q)$

$\text{qnorm}(q) = \Phi^{-1}(q) = F^{-1}(q)$

$x = \mu + \sigma z$ qnorm(q, μ, σ) Another variant (inverse distribution).

MATH 104 L03

2.4 Fitting Distributions to Data

Method of Moments -

A moment measures / characterises the shape of the probability distribution. The first moment is the mean of the distribution. In general, the k^{th} moment =

$$\mu_k = E(X^k)$$

Sample Moments -

Sample moments are an estimate of the population moments. The k^{th} sample moment for data $\{x_i\}_{i=1}^{n}$ is defined:

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

Population Moments -

$$E(X^k; \hat{\theta}) = \frac{1}{n} \sum_{i=1}^{n} x_i^k$$

this is an estimator $\hat{\theta}$ of a set of parameters θ.

Bernoulli Method of Moments -

$$E(X; \hat{\theta}) = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\Rightarrow \hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Geometric Method of Moments -

$$E(X; \hat{\theta}) = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\Rightarrow \frac{1-\hat{\theta}}{\hat{\theta}} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\Rightarrow \hat{\theta} = \frac{n}{n + \sum_{i=1}^{n} x_i}$$

n = number of successes

$n + \sum_{i=1}^{n} x_i$ = total number of attempts

Poisson Method of Moments -

$$E(X; \hat{\theta}) = 1/n \sum_{i=1}^{n} x_i$$

$$\Rightarrow \hat{\theta} = 1/n \sum_{i=1}^{n} x_i$$

Uniform Method of Moments -

$\theta = (a, b)$ Mean $= \frac{(a+b)}{2}$ Variance $= \frac{(a-b)^2}{12}$

$$E(X; \hat{\theta}) = 1/n \sum_{i=1}^{n} x_i$$

Simultaneous Equations

$$E(X^2; \hat{\theta}) = 1/n \sum_{i=1}^{n} x_i^2$$

$$\hat{a} = 2/n \sum_{i=1}^{n} x_i - \hat{b} = \bar{x} - \sqrt{\frac{3(n-1)}{n}}\, s(x)$$

Normal Method of Moments -

$$E(X; \hat{\theta}) = 1/n \sum_{i=1}^{n} x_i$$

same Simultaneous Equations

$$\Rightarrow \hat{\mu} = 1/n \sum_{i=1}^{n} x_i$$

$$E(X^2; \hat{\theta}) = 1/n \sum_{i=1}^{n} x_i^2$$

$$\Rightarrow \hat{\sigma}^2 + \hat{\mu}^2 = 1/n \sum_{i=1}^{n} x_i^2$$

$$\hat{\sigma}^2 + \hat{\mu}^2 = 1/n \sum_{i=1}^{n} x_i^2$$

$$\Rightarrow \hat{\sigma}^2 = \frac{(n-1)}{n} s^2(x)$$

Q-Q Plots -

There are two visual methods for checking the assumption of normality. The first is a histogram overlaid by a normal curve and the second is examining a Q-Q plot. The closer the points are to a straight line the more they fit the specified model. These are calculated using R.

3/3/2022

MATH 104 L04

Point Estimates -

The most intuitive way to estimate the population mean based on a sample is to calculate the sample mean $\bar{x} = 1/n \sum_{i=1}^{n} x_i$ point estimate of the population mean.

Point estimates aren't exact and so a running mean is a sequence of means where each mean gets more accurate each iteration as it uses data before it.

Standard Error of the Mean -

If we sample for the mean amount of time multiple times we can create a sampling distribution for the sample mean. This represents the distribution of the point estimates based on samples of a fixed size. The standard deviation of the sample mean describes the typical error of the point estimate and is called the standard error of an estimate. $=$ or $\frac{\sigma \bar{x}}{\sqrt{n}}$

e.g. where $\sigma = 4.9$ over 100 samples SEM $= \frac{4.9}{10}$

MATH 104 LO5

Confidence Interval -

This is a plausible range of values for the population parameter from the data sample. Here we focus on building a confidence interval for the population mean.

e.g 95% confidence interval:

$$\bar{x} \pm 1.96 \frac{s}{\sqrt{n}}$$

We assume independence (This is satisfied if we have a random sample < 10 % of the population).

An x% confidence interval gives us x% confidence that the population parameter is inside the interval.

90% → qnorm(0.95)

95% → qnorm(0.975)

99% → qnorm(0.995)

MATH 104 LOG

Hypothesis Testing

We have a null and alternative hypothesis. The null hypothesis is to be disproved or not.

	Test Conclusion: Don't reject H_0	Reject H_0 for H_1
H_0 true	✓	Type 1 Error
H_1 true	Type 2 Error	✓

We calculate a p-value, set a significance level and if $p < \alpha$ (significance level) reject H_0, if not keep H_0.

Calculating p-values

1) We need μ_0 (null value), sample size (n), sample mean ($\bar{x}$), sample standard deviation (s)
2) Calculate SEM $s/\sqrt{n}$
3) Calculate z-score of sample mean $Z = \dfrac{\bar{x} - \mu_0}{SEM}$
4) use R Studio and calculate pnorm(Z)
5) Find p-value

$H_1: \mu < \mu_0$ or $H_1: \mu > \mu_0$ or $H_1: \mu \neq \mu_0$

$p = \text{pnorm}(Z)$ ~~p = pnorm(z)~~ $p = 1 - \text{pnorm}(Z)$ $p = 2\,\text{pnorm}(-|Z|)$

A p-value is the probability of observing something as extreme or more extreme assuming the null hypothesis is true.

MATH 104 L07

Central Limit Theorem -

If $X_1, \dots, X_n$ are independent and identically distributed random variables with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$ then,

$$\sum_{i=1}^{n} X_i \sim N(n\mu, n\sigma^2) \qquad \bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

approximately (as $n \to \infty$) irrespective of the original distribution of X_i.

The distribution of the sum or average of n independent and identically distributed random variables approaches a normal as n gets large.

MATH 104 L08

3.1 Paired Data -

When two sets of observations have a correspondence they are said to be paired. To analyse paired data we look at the outcomes of each pair of observations.

We compute standard error associated with $\bar{x}_{diff}$ using standard deviation of the differences and number of differences

$$SE_{\bar{x}_{diff}} = \frac{s_{diff}}{\sqrt{n_{diff}}}.$$

To find the interval identify z^* and plug it, the point estimate and the standard error into the confidence interval formula.

$$= \text{point estimate} \pm z^* SE$$

3.2 Difference of Two Means -

Difference in two population means $\mu_1 - \mu_2$ given that the data are not paired.

1) Identify conditions to ensure a point estimate of the difference $\bar{x}_1 - \bar{x}_2$ is nearly normal
2) Introduce a formula for the standard error.

Conditions for normality -

If the sample means $\bar{x}_1$ and $\bar{x}_2$ each meet the criteria for having nearly normal sampling distributions and the observations in the two samples are independent then the difference in sample means $\bar{x}_1 - \bar{x}_2$ will have a sampling distribution that is nearly normal.

$$SE_{\bar{x}_w - \bar{x}_m} = \sqrt{\frac{\sigma_w^2}{n_w} + \frac{\sigma_m^2}{n_m}}$$

Distribution of a difference of sample means.

24/3/2022

MATH 104 L09

3.3 One-sample means with the t distribution -

Central Limit Theorem for normal data -

The sampling distribution of the mean is nearly normal when the sample observations are independent and come from a nearly normal distribution. This is true for any sample size.

The t distribution -

A t distribution has a bell shape. However it's tails are thicker than the normal's models. This means observations are more likely to fall beyond two standard deviations from the mean. These extra thick tails are exactly the correction we need to resolve the problem of a poorly estimated standard error.

The t distribution always centred at zero has a single parameter: degrees of freedom. They describe the precise form of the bell-shaped t distribution.

When the degrees of freedom is about 30 or more the t distribution is nearly indistinguishable from the normal.

t distribution as a solution to standard error -

We must check two conditions -

* Independence of observations. Collect a random sample <10% of population
* Observations come from a nearly normal distribution. We often i) take a look at a plot of the data for obvious departures and ii) consider whether any previous experiences alert us the data may not be nearly normal.

One sample t confidence intervals -

$\bar{x} \pm t^*_{df} SE$ to determine width of confidence interval.

Degrees of freedom for a single sample - If the sample has n observations and we are examining a single mean then we use the t distribution with $df = n-1$ degrees of freedom.

One sample t tests -

We standardise the sample mean.

$$T = \frac{\bar{x} - \text{null value}}{SE}$$

Two sample t test -

Test statistic

$$T = \frac{\text{point estimate} - \text{null value}}{SE}$$

24/3/2022

MATH 104 L10

4.1 Inference for a single proportion -

A sample peroportion can be described as a sample mean. If we represent a success as 1 and fail as 0. Sample proportion is mean of these outcomes.

$$\hat{p} = \frac{0+1+1+\dots+0}{976} \quad \text{for example } (976) = n$$

Conditions for $\hat{p}$ being nearly normal.

a) sample observations are independent and

b) we expected to see at least 10 successes and fails in the sample. $np \geq 10$ and $n(1-p) \geq 10$

This is the success - failure condition.

If these conditions are met:

$$SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

Hypothesis testing for a proportion -

null value (single proportion) -

$$SE = \sqrt{\frac{p_0(1-p_0)}{n}}$$

$$Z = \frac{\text{point estimate} - \text{null value}}{SE}$$

Difference of two proportions -

Conditions for $\hat{p_1} - \hat{p_2}$ to be normal.

a) Each proportion separately follows a normal model

b) Two samples are independent of each other.

$$SE_{\hat{p_1}-\hat{p_2}} = \sqrt{SE_{\hat{p_1}}^2 + SE_{\hat{p_2}}^2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Hypothesis testing when $H_0: p_1 = p_2$

We can estimate

$$\hat{p} = \frac{\text{\# of successes}}{\text{\# of cases}}$$

This is called the pooled estimate and is used to compute the standard error when $H_0: p_1 = p_2$.

$$\hat{p_1} = \frac{\text{\# of successes in sample 1}}{n_1}$$

$$Z = \frac{\text{point estimate - null value}}{SE}$$