6.S085 Statistics for Research Projects **IAP 2014**

Lecture 2: January 22

Lecturer: Ramesh Sridharan and George Chen *Notes by: William Li*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

2.1 Confidence Intervals

French baker example:

- Gaussian distribution around 950g
- skewed distribution around 1000g

Goal: estimate the parameters of a binomial random variable

Setup: $x_1, x_2, ..., x_n$

Data points are binary

There is some true distributions ("population distribution") with true parameters p

Use data to draw conclusions about p

How to estimate p?

$\hat{p} = \frac{1}{n} \sum_i x_i$

p isn't random, but x_i are random, so $\hat{p}$ is random!

$$\begin{aligned} E[\hat{p}] &= E[\frac{1}{n} \sum x_i] && (2.1) \\ &= \frac{1}{n} \sum E[x_i] && (2.2) \\ &= p && (2.3) \end{aligned}$$

$$\begin{aligned} var[\hat{p}] &= var[\frac{1}{n} \sum x_i] && (2.4) \\ &= \frac{1}{n^2} \sum var[x_i] && (2.5) \\ &= \frac{1}{n^2} np(1-p) && (2.6) \\ &= \frac{p(1-p)}{n} && (2.7) \end{aligned}$$

The variance on $\hat{p}$ goes down with more n.

By central limit therorem, $\hat{p}$ is approximately Gaussian: Completely characterized by mean and variance:

$\hat{p}\ N(p, \frac{p(1-p)}{n}$

Confidence intervals: loosely, we want "the range of intervals where p probably is"

If we repeat sampling (data collection) and CI computation, 95% of those repeats will give an interva that has p

Recall that, in a Gaussian, 95% of the data lies within 2 standard deviations of the mean

$\hat{p} \pm 2\sqrt{\frac{p(1-p)}{n}}$: This is our 95% confidence interval

What a CI means:

Probability of getting a $\hat{p}$ within 2 standard deviations of mean p

Since we don't actually know p, we will replace p with $\hat{p}$!

$\hat{p} \pm 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$: This is our 95% confidence interval

"With probability 0.95, $\hat{p}$ will just such that this confidence intervals contains p

"accurate to within $2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ percentage points, 95 times out of 100."

Note that, with more n, we will have a tighter confidence interval

2.2 Hypothesis Testing

We have a hypothesis about p

Use data to see whether we can reject the hypothesis

Find out how likely the data is under the hypothesis; if it's very unlikely, reject

Null and alternative hypothesis formulation:

$H_0 : p = p^*$ (null hypothesis)

$H_a : p > p^*, p < p^*, p \neq p^*$ (alternative hypothesis)

One-tailed test: $>, <$: only interested in one direction

Two-tailed test: $\neq$: interested in both or either side

Tail of Gaussian represented by $alpha$

Type I error/false positive: reject H_0 when H_0 is true

Type II error/false negative: fail to reject H_0 when H_0 is false

A hypothesis test can only tell you whether you can accept the null hypothesis or fail to accept the null hypothesis

$\alpha = 0.05$: Probability of false positive

"There is only a 5% chance that the observation happened by chance if the null observation is true"

"What is the probability of getting this value if the null hypothesis were true?"

2.2.1 p values

We have a null hypothesis value p^*, and then we observe x

We ask: "What is the probability of getting x (or something more extreme) if null hypothesis is true?"

This is what a p value is

2.2.2 Statistical Power

$1 - P(\text{Type II errror})$

Probability of rejecting H_0 when it's wrong

Defined with respect to a particular alternative value

Null hypothesis is centered around a particular value

The power depends on the threshold and α

Distributions that are closer together: reduces the power of the test

"If the true value is 0.4, power is the shaded area"

2.3 Continuous Hypothesis Testing, Confidence Intervals

observe $x_1, x_2, ..., x_n$ (comes from Gaussian distribution with mean μ, true variance σ^2

Start with the (unrealistic) assumption that σ is known

Want to estimate μ

$\hat{\mu} = \frac{1}{n} \sum x_i$

$E[\hat{\mu}] = \mu$

$$var\hat{\mu} = \frac{1}{n^2} \sum var[x_i] \tag{2.8}$$

$$\frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n} \tag{2.9}$$

$\frac{\sigma}{\sqrt{n}}$ is the standard error fo the mean

In general, stnadard error of a statistic is the standard deviation of its sampling distribution

$Z = \frac{\hat{\mu} - \mu}{\sigma / sqrtn}$

$Z\ N(0, 1)$ ("test statistic")

2.3.1 Hypothesis Testing

Hypothesis testing: μ is known; it's the null distance's mean

2.3.2 Confidence Intervals

$P(-2 \leq z \leq 2)$

$P(-2 \leq \frac{m\hat{u}-\mu}{sigma/\sqrt{n}} \leq 2)$

$P(\hat{\mu} - 2\frac{\sigma}{\sqrt{n}} \leq \mu \leq \hat{\mu} + 2\frac{\sigma}{\sqrt{n}})$

CI: $\hat{\mu} \pm 2\frac{\sigma}{\sqrt{n}}$

What if we don't know σ?

Use:

$\sigma^2 = \frac{1}{n-1}\sum(x_i - \hat{\mu}^2\ \sigma^2 = s^2$

Approximate standard error: $s/\sqrt{n}$ / $t = \frac{\hat{\mu}-\mu}{s/\sqrt{n}}$

t has a t distribution with $n-1$ degrees of freedom

Now we want $P(-X \leq t \leq +X) = 0.95$

Aside: $\frac{(n-1)}{\sigma^2}s^2\ \chi^2$ with n-1 degrees of freedom

Student t distribution

$Z\ N(0,1)$

$\mu\ \chi^2$ with r degrees of freedom

$t = \frac{z}{\sqrt{u/r}}$ has Student t distribution

2.4 Two-Sample Tests

$x_1, ..., x_n$

$y_1, ..., y_n$

Matched pairs (correspondance), i.e. $x_1 \rightarrow y_1$, $x_2 \rightarrow y_2$

Then we can define $w_i = y_i - x_i$ run a one-sample test on w

2.4.1 Pooled Variance

"homoskedastic" − variances are the same

$s^2 = \frac{(n-1)s_x^2+(m-1)s_y^2}{n+m-2}$

2.4.2 Unpooled Variances

; "heteroskedastic" – variances are not the same

Test statistic no longer exactly t-distribution

2.5 Warnings About Tests

For hypothesis tests:

- Don't say anything stronger than "the data do not support the null hypothesis"
- Avoid multiple comparisons, or correct (Bonferrnoni correction, divides p-value by the number of comparisons

False discovery rate (FDR)