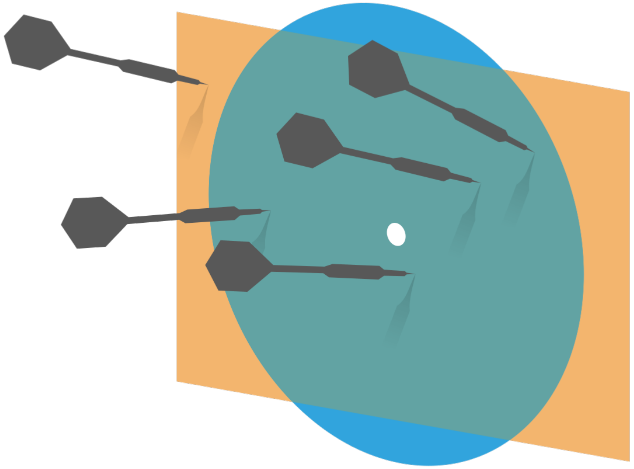
Probability and Statistics for Computer Science



Credit: wikipedia

"Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write." H. G. Wells

Hongye Liu, Teaching Assistant Prof, CS361, UIUC, 10.15.2020

Last Lecture

- Review of sample mean and confidence interval
- Bootstrap simulation of sample statistic A tale of two statisticians
- Hypothesis test intro

Objectives

- Hypothesis test of popmean
- Chi-square test
- Maximum Likelihood Estimation

parameter inference

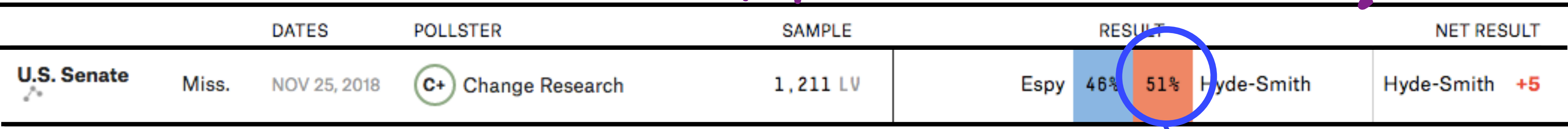
A hypothesis

- Ms. Smith’s vote percentage is 55%

  This is what we want to test, often called null hypothesis H_0

H_0 : popmean = 55%

H_1 : popmean ≠ 55%



		DATES	POLLSTER	SAMPLE	RESULT				NET RESULT
U.S. Senate	Miss.	NOV 25, 2018	C+ Change Research	1,211 LV	Espy	46%	51%	Hyde-Smith	Hyde-Smith +5

One realized Sample mean : 51%

- Should we reject this hypothesis given the poll data?

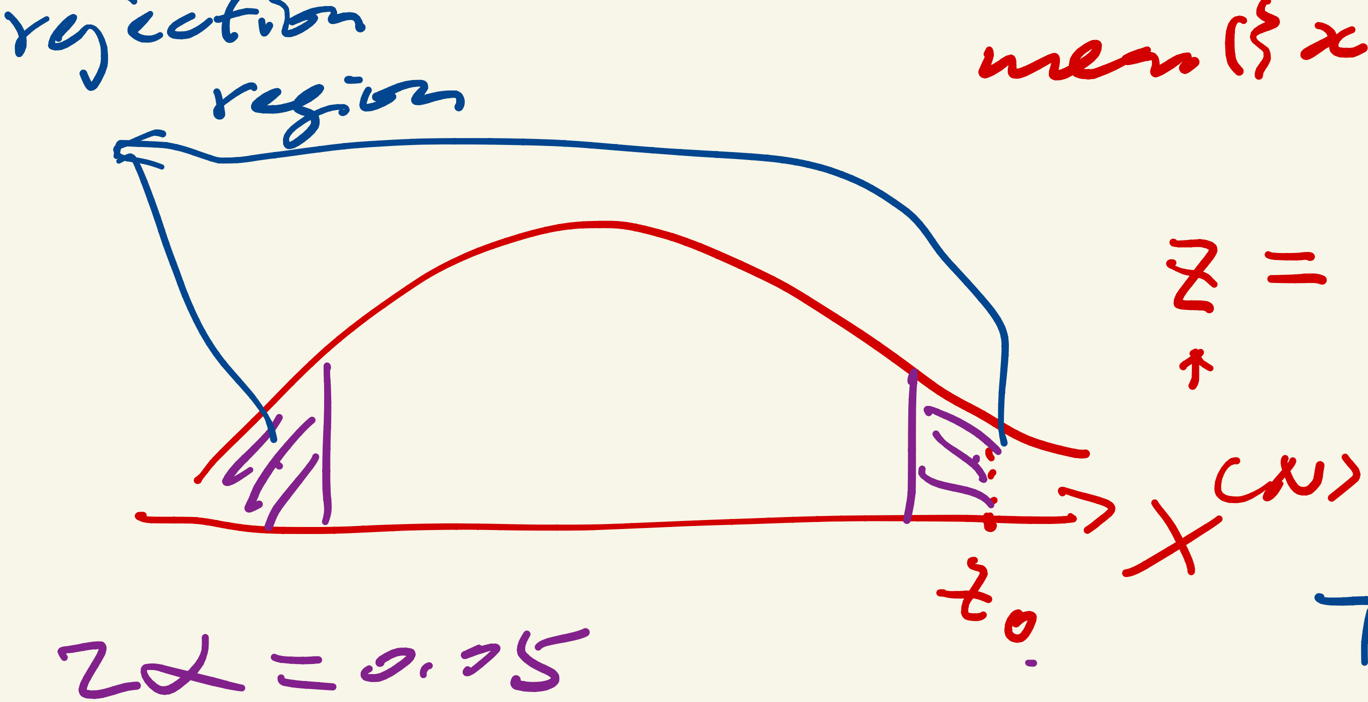
H_0 : popmean({X}) = v_0

H_1 : popmean({X}) $\neq$ v_0

{x} sample mean value

mean({x})

$N > 30$



$$Z = \frac{\text{mean}(\{x\}) - \underline{\text{popmean}}_{\,v_0}}{\text{stderr}(\{x\})}$$

$2\alpha = 0.05$

$$T = \frac{\text{mean}(\{x\}) - \text{popmean}}{\text{stderr}(\{x\})}$$

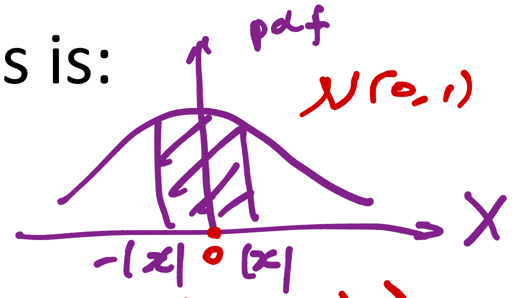
Fraction of “less extreme” samples

- Assuming the hypothesis H_0 is true
- Define a test statistic

$$x = \frac{\overset{\text{realized}}{(sample\ mean)} - \overset{\nearrow\ 55\%}{(hypothesized\ value)}}{\underset{\text{realized}}{standard\ error}}$$

- Since *N*>30, x should come from a standard normal
- So, the fraction of “less extreme” samples is:

$$f = \frac{1}{\sqrt{2\pi}} \int_{-|x|}^{|x|} exp(-\frac{u^2}{2}) du$$



$2(CDF(|x|-0.5))$

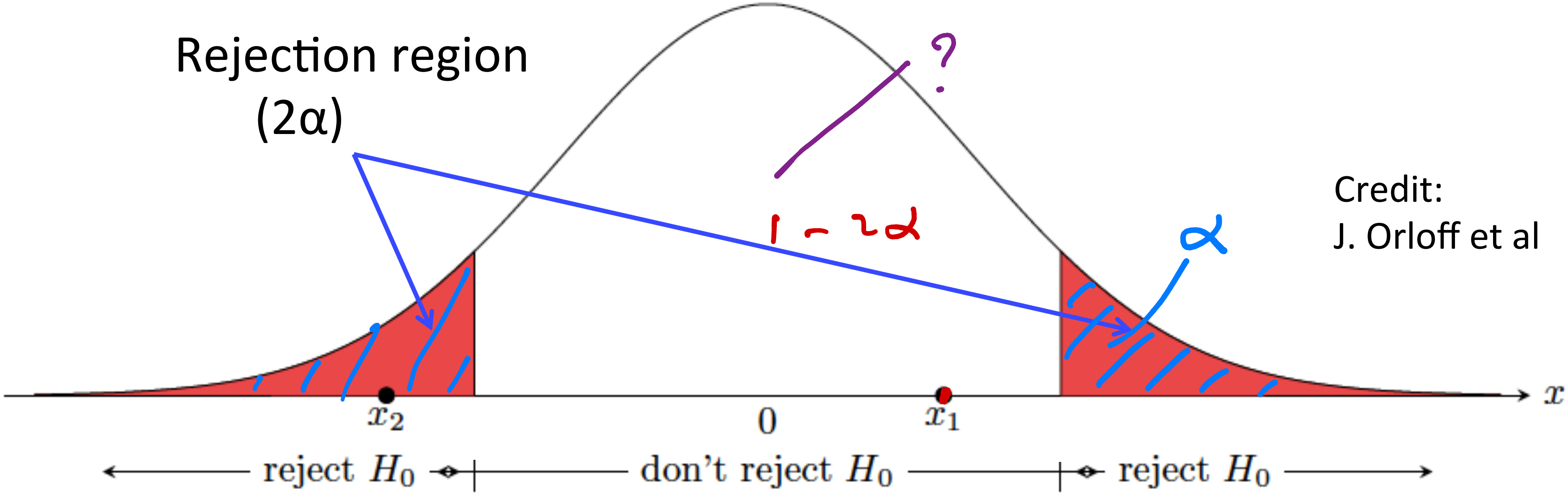
$2CDF(|x|)-1$

Rejection region of null hypothesis H_o

- Assuming the hypothesis H_0 is true
- Define a test statistic

$$x = \frac{(sample\ mean) - (hypothesized\ value)}{standard\ error}$$

- Since *N*>30, x should come from a standard normal



Credit:
J. Orloff et al

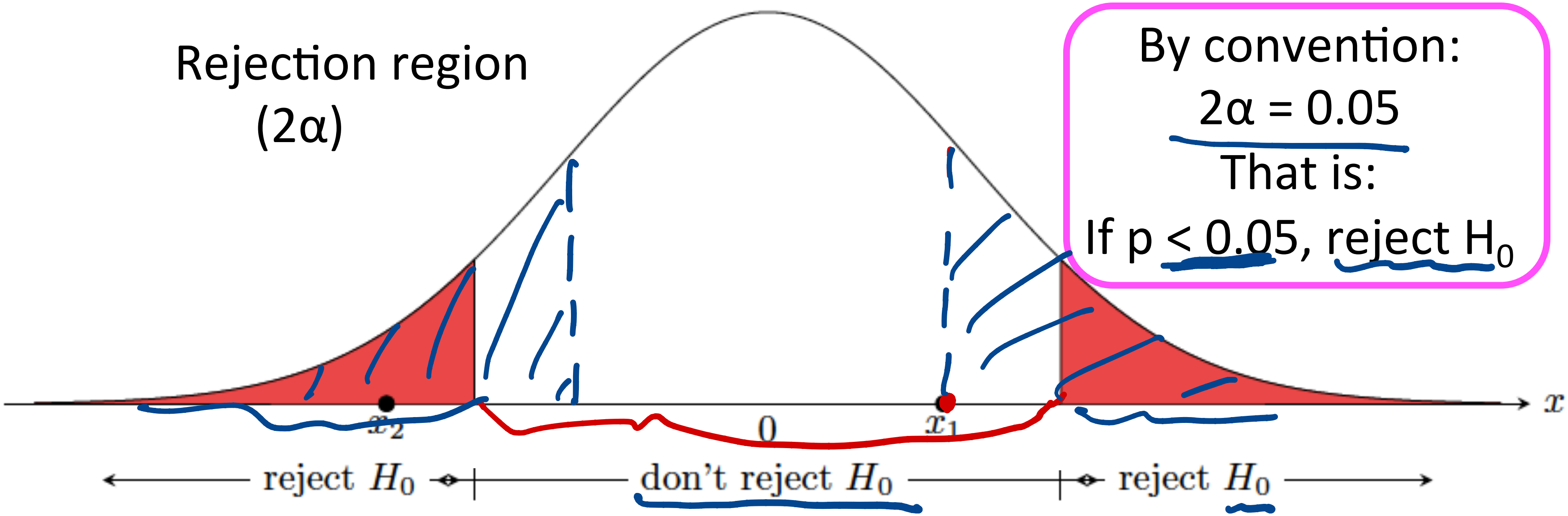
P-value: Rejection region- “The extreme fraction”

- It is conventional to report the p-value of a hypothesis test

$$p = 1 - f = 1 - \frac{1}{\sqrt{2\pi}} \int_{-|x|}^{|x|} exp(-\frac{u^2}{2})du$$

0.05

- Since *N*>30, x should come from a standard normal



Rejection region (2α)

By convention:

2α = 0.05

That is:

If $p \leq 0.05$, reject H_0

p-value: election polling

- $H_{0:}$ Ms. Smith's vote percentage is 55%
- The sample mean is 51% and stderr is 1.44%
- The test statistic $x = \frac{51 - 55}{1.44} = -2.7778$
- And the p-value for the test is:

$$p = 1 - \frac{1}{\sqrt{2\pi}} \int_{-2.7778}^{2.7778} exp(-\frac{u^2}{2})du = 0.00547 \qquad < 0.05$$

- So we reject the hypothesis

Hypothesis test if N < 30

- Q: what distribution should we use to test the hypothesis of sample mean if N<30?

A. Normal distribution

B. t-distribution with degree =30

C. t-distribution with degree = N

[✓ D.] t-distribution with degree = N-1

The use and misuse of p-value

- p-value use in scientific practice
  - Usually used to reject the null hypothesis that the data is random noise
  - Common practice is $p < 0.05$ is considered significant evidence for something interesting

- Caution about p-value hacking
  - Rejecting the null hypothesis doesn't mean the alternative is true
  - $P < 0.05$ is arbitrary and often is not enough for controlling false positive phenomenon

Be wary of one tailed p-values

- The one tailed p-value should only be considered when the realized sample mean or differences will for sure fall only to one size of the distribution.

- Sometimes scientist are tempted to use one tailed test because it'll give smaller p-val. But this is bad statistics!

Chi-square distribution

- If $Z_i's$ are independent variables of standard normal distribution, $X = Z_1^2 + Z_2^2 + \ldots + Z_m^2 = \sum_{i=1}^{m} Z_i^2$

  has a Chi-square distribution with degree of freedom m, $X \sim \chi^2(m)$

- We can test the goodness of fit for a model using a statistic ***C*** against this distribution, where

  observed theoretical

$$C = \sum_{i=1}^{m} \frac{(\underline{f_o}(\varepsilon_i) - \underline{f_t}(\varepsilon_i))^2}{f_t(\varepsilon_i)}$$

Independence analysis using Chi-square

- Given the two way table, test whether the column and row are independent

Elementary school students data

observed data

	Boy	Girl	Total
Grades	117	130	247
Popular	50	91	141
Sports	60	30	90
Total	227	251	478

"Popular Kids"

$\frac{227}{251}$

Independence analysis using Chi-square

- The theoretical expected values if independent

$$C = \sum_{i=1}^{?\,2} \frac{(f_o - f_t)^2}{f_t}$$

r

	Boy	Girl	Total	
Grades	117.29916	129.70084	247	✓
Popular	66.96025	74.03975	141	✓
Sports	42.74059	47.25941	90	✓
Total	227	251	478	
	✓	✓		

$6 - 5 + 1$

$= 2$

The degree of the chi-square distribution for the two way table

- The degree of freedom for the chi-square distribution for a ***r*** by ***c*** table is

  $\underline{(r-1) \times (c-1)}$ where r>1 and c>1

- Because the degree df = n-1-p See textbook Pg 171-172

$$= rc - 1 - (r-1) - (c-1)$$

n is the number of cells of data;
p is the number of unknown parameters

$$= (r-1) \times (c-1)$$

$$= 2$$

Chi-square test for the popular kid data

- The Chi-statistic : 21.455

```
chisq.test(data_BG)

	Pearson's Chi-squared test

data:  data_BG
X-squared = 21.455, df = 2, p-value = 2.193e-05
```

- P-value: 2.193e-05
- It’s very unlikely the two categories are independent

Q. What is the degree of freedom for this?

- The following 2-way table for chi-square test has a degree of freedom equal to: (4-1)(5-1)

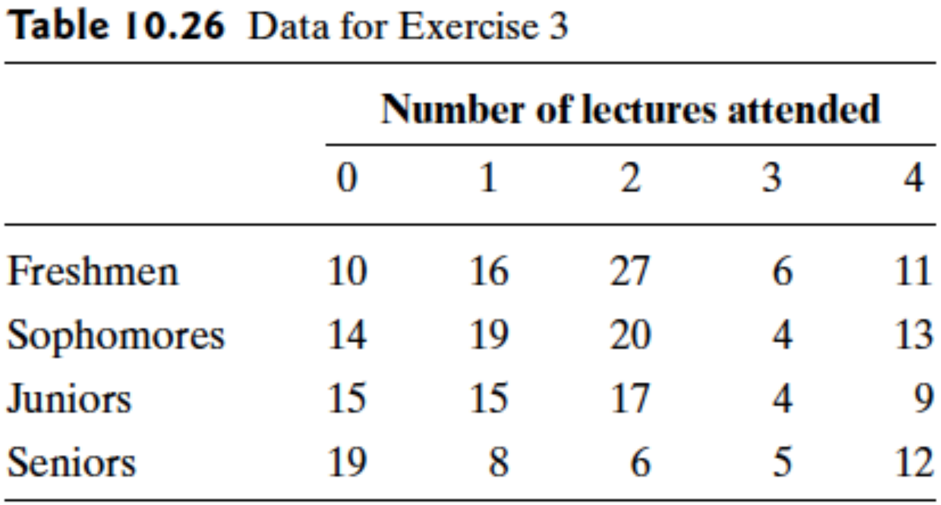


Table 10.26 Data for Exercise 3

	Number of lectures attended				
	0	1	2	3	4
Freshmen	10	16	27	6	11
Sophomores	14	19	20	4	13
Juniors	15	15	17	4	9
Seniors	19	8	6	5	12

ref → Degroot et al.
pg 652

A. 20 B. 9

[✓ C.] 12 D. 4

Chi-square test is very versatile

- Chi-square test is so versatile that it can be utilized in many ways either for discrete data or continuous data via intervals

- Please check out the worked-out examples in the textbook and read more about its applications.

Maximum likelihood estimation

$$\underline{P}(X=k) = \binom{N}{k} \underset{\uparrow}{p}^{k} (1-p)^{N-k}$$

$p \rightarrow$ unknown

write p as $\underline{\theta}$

$$= \underline{\binom{N}{k} \theta^{k} (1-\theta)^{N-k}}$$

Maximize $P(X=k)$ give $\hat{\theta}$

$$\hat{\theta} = \underset{\theta}{\mathrm{Argmax}}\ P(\theta)$$

The parameter estimation problem

- Suppose we have a dataset that we know comes from a distribution (ie. Binomial, Geometric, or Poisson, etc.)

- What is the best estimate of the parameters (**θ** or **θ**s) of the distribution?

- Examples:
  - For binomial and geometric distribution, **θ** = *p* (probability of success)
  - For Poisson and exponential distributions, **θ** = λ (intensity)
  - For normal distributions, **θ** could be $\boldsymbol{\mu}$ or $\boldsymbol{\sigma^2}$.

$$\underset{k \geq 0}{P(X=k)} = \frac{e^{-\lambda} \lambda^k}{k!}$$

Motivation: Poisson example

- Suppose we have data on the number of babies born each hour in a large hospital

hour	**1**	**2**	**…**	*N*
# of babies	k_1	k_2	**…**	k_N

- We can assume the data comes from a Poisson distribution

- What is your best estimate of the intensity λ?

Credit: David Varodayan

Maximum likelihood estimation (MLE)

- We write the probability of seeing the data D given parameter θ

$$\underline{L(\theta) = \binom{N}{K} \theta^{K} (1-\theta)^{N-K}}$$

$$\boxed{\underline{L(\theta) = P(D|\theta)}}$$

$D: K, N$

- The **likelihood function** $L(\theta)$ is **not** a probability distribution

regarding θ

- The **maximum likelihood estimate (MLE)** of θ is

$$\boxed{\underset{\uparrow}{\hat{\theta}} = arg \max_{\theta} L(\theta)}$$

$\int L(\theta) \neq 1$

$\sum L(\theta) \neq 1$

Why is $L(\theta)$ not a probability distribution?

A. It doesn’t give the probability of all the possible θ values.

B. Don’t know whether the sum or integral of $L(\theta)$ for all possible θ values is one or not.

[✓ C.] Both.

Likelihood function: Binomial example

- Suppose we have a coin with unknown probability of coming up heads

- We toss it ***N*** times and observe ***k*** heads

- We know that this data comes from a binomial distribution

- What is the likelihood function $L(\theta) = P(D|\theta)$?

Likelihood function: binomial example

- Suppose we have a coin with unknown probability of coming up heads

- We toss it ***N*** times and observe ***k*** heads

- We know that this data comes from a binomial distribution

- What is the likelihood function $L(\theta) = P(D|\theta)$?

$$\boxed{L(\theta) = \binom{N}{k} \theta^k (1-\theta)^{N-k}}$$

$\theta = p$

unknown

MLE derivation: binomial example

$$L(\theta) = \binom{N}{k}\theta^k(1-\theta)^{N-k}$$

In order to find: $\hat{\theta} = arg\ \underset{\theta}{max}\ L(\theta)$

We set: $$\frac{\mathrm{d}L(\theta)}{\mathrm{d}\theta} = 0$$

MLE derivation: binomial example

$$L(\theta) = \binom{N}{k} \theta^k (1-\theta)^{N-k}$$

MLE derivation: binomial example

$$\boxed{L(\theta) = \binom{N}{k}\theta^k(1-\theta)^{N-k}}$$

$$\frac{d}{d\theta}L(\theta) = \binom{N}{k}(k\theta^{k-1}(1-\theta)^{N-k} - \theta^k(N-k)(1-\theta)^{N-k-1}) = 0$$

MLE derivation: binomial example

$$\boxed{L(\theta) = \binom{N}{k}\theta^k(1-\theta)^{N-k}}$$

$$\frac{d}{d\theta}L(\theta) = \binom{N}{k}(k\theta^{k-1}(1-\theta)^{N-k} - \theta^k(N-k)(1-\theta)^{N-k-1}) = 0$$

$$k\theta^{k-1}(1-\theta)^{N-k} = \theta^k(N-k)(1-\theta)^{N-k-1}$$

MLE derivation: binomial example

$$\boxed{L(\theta) = \binom{N}{k} \theta^k (1-\theta)^{N-k}}$$

$$\frac{d}{d\theta} L(\theta) = \binom{N}{k} (k\theta^{k-1}(1-\theta)^{N-k} - \theta^k (N-k)(1-\theta)^{N-k-1}) = 0$$

$$k\theta^{k-1}(1-\theta)^{N-k} = \theta^k (N-k)(1-\theta)^{N-k-1}$$

$$k - k\theta = N\theta - k\theta$$

MLE derivation: binomial example

$$\boxed{L(\theta) = \binom{N}{k}\theta^k(1-\theta)^{N-k}}$$

$$\theta = \frac{1}{m}$$

$$\binom{N}{k}\left(\frac{1}{m}\right)^{k}\left(1-\frac{1}{m}\right)^{N-k}$$

$$\frac{d}{d\theta}L(\theta) = \binom{N}{k}(k\theta^{k-1}(1-\theta)^{N-k} - \theta^k(N-k)(1-\theta)^{N-k-1}) = 0$$

$$\alpha = m$$

$$k\theta^{k-1}(1-\theta)^{N-k} = \theta^k(N-k)(1-\theta)^{N-k-1}$$

$$k - k\theta = N\theta - k\theta$$

$$L(\alpha) = \ldots$$

$$\boxed{\hat{\theta} = \frac{k}{N}}$$

The MLE of p

Likelihood function: geometric example

- Suppose we have a die with unknown probability of coming up six

- We roll it and it comes up six for the first time on the kth roll

- We know that this data comes from a geometric distribution

- What is the likelihood function $L(\theta) = P(D|\theta)$? **Assume θ is p**.

MLE derivation: geometric example

$$\boxed{L(\theta) = (1-\theta)^{k-1}\theta}$$

MLE derivation: geometric example

$$\boxed{L(\theta) = (1-\theta)^{k-1}\theta}$$

$$\frac{d}{d\theta}L(\theta) = (1-\theta)^{k-1} - (k-1)(1-\theta)^{k-2}\theta = 0$$

$$(1-\theta)^{k-1} = (k-1)(1-\theta)^{k-2}\theta$$

$$1-\theta = k\theta - \theta$$

$$\hat{\theta} = \frac{1}{k}$$

The MLE of p

MLE with data from IID trials

- If the dataset $D = \{x\}$ comes from IID trials

$$L(\theta) = P(D|\theta) = \prod_{x_i \in D} P(x_i|\theta)$$

- Each x_i is one observed result from an IID trial

K_1 1st hr

K_2 2nd hr

$\vdots$

K_m mth hr

Q: MLE with data from IID trials

- If the dataset $D = \{x\}$ comes from IID trials

$$L(\theta) = P(D|\theta) = \prod_{x_i \in D} P(x_i|\theta)$$

$D : \{x_1, x_2, \ldots x_m\}$

- Why is the above function defined by the product?

A. IID samples are independent

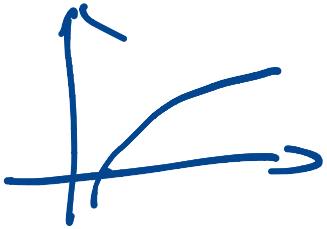
B. Each trial has identical probability function

[✓ C.] Both.

MLE with data from IID trials

- If the dataset $D = \{x\}$ comes from IID trials

$$L(\theta) = P(D|\theta) = \prod_{x_i \in D} P(x_i|\theta)$$



- The likelihood function is hard to differentiate in general, except for the binomial and geometric cases.

$$\log(A \cdot B)$$
$$= \log A + \log B$$

- Clever trick: take the (natural) log

Log-likelihood function

- Since log is a strictly increasing function

$$\hat{\theta} = arg\ \max_{\theta}\ L(\theta) = arg\ \max_{\theta}\ logL(\theta)$$

- So we can aim to maximize the **log-likelihood function**

$$logL(\theta) = logP(D|\theta) = log \prod_{x_i \in D} P(x_i|\theta) = \sum_{x_i \in D} logP(x_i|\theta)$$

- The log-likelihood function is usually much easier to differentiate

Log-likelihood function: Poisson example

- Suppose we have data on the number of babies born each hour in a large hospital

hour	**1**	**2**	**…**	*N*
# of babies	k_1	k_2	**…**	k_N

- We can assume the data comes from a Poisson distribution λ

- What is the log likelihood function $LogL(\theta)$?

Log-likelihood function: Poisson example

$$L(\theta) = \prod_{i=1}^{N} \boxed{\frac{e^{-\theta}\theta^{k_i}}{k_i!}}$$

$$\frac{e^{-\lambda \to \theta}\,\lambda^{k_i}}{k_i!} \qquad \lambda \to \theta$$

$$log\ L(\theta) = log\ (\prod_{i=1}^{N} \frac{e^{-\theta}\theta^{k_i}}{k_i!}) = \sum_{i=1}^{N} log(\frac{e^{-\theta}\theta^{k_i}}{k_i!})$$

$$= \sum_{i=1}^{N} (-\theta + k_i\ log\theta - log\ k_i!)$$

MLE : Poisson example

$$LogL(\theta) = \sum_{i=1}^{N}\underline{(-\theta + k_i\ log\theta - log\ k_i!)}$$

$$\uparrow$$

$$\frac{d\,LogL(\theta)}{d\theta} = \sum_{i=1}^{N}\left(-1 + k_i\frac{1}{\theta} - 0\right)$$

$$= \sum_{i=1}^{N}\left(-1 + \frac{k_i}{\theta}\right) = 0$$

$$\hat{\theta}$$

MLE : Poisson example

$$LogL(\theta) = \sum_{i=1}^{N} (-\theta + k_i \ log\theta - log \ k_i!)$$

$$\frac{d}{d\theta} log \ L(\theta) = 0 \ \Rightarrow \sum_{i=1}^{N} (-1 + \frac{k_i}{\theta} - 0) = 0$$

MLE : Poisson example

$$LogL(\theta) = \sum_{i=1}^{N}(-\theta + k_i \ log\theta - log \ k_i!)$$

$$\frac{d}{d\theta} log \ L(\theta) = 0 \Rightarrow \boxed{\sum_{i=1}^{N}(-1} + \boxed{\frac{k_i}{\theta}} - 0) = 0$$

$$\underline{-N} + \frac{\sum_i^N k_i}{\theta} = 0$$

MLE : Poisson example

$$LogL(\theta) = \sum_{i=1}^{N}(-\theta + k_i\ log\theta - log\ k_i!)$$

$$\frac{d}{d\theta}log\ L(\theta) = 0 \Rightarrow \sum_{i=1}^{N}(-1 + \frac{k_i}{\theta} - 0) = 0$$

$$-N + \frac{\sum_i^N k_i}{\theta} = 0$$

$$\boxed{\hat{\theta} = \frac{\sum_i^N k_i}{N}}$$

The MLE of λ

MLE for normal distribution

- Suppose we model the dataset $D = \{x\}$ as normally distributed

- What should be the likelihood function? Is the method of modeling the same as for the Poisson distribution?

A. Yes B. No

pdf for $L(\theta)$

MLE for normal distribution

- Suppose we model the dataset $D = \{x\}$ as normally distributed

- What should be the likelihood function? Is the method of modeling the same as for the Poisson distribution? **Yes and No**. The idea is similar but the normal distribution is continuous, we need to use the **probability density** instead.

$\mu \quad \sigma^2$

MLE for normal distribution

- Suppose we model the dataset $D = \{x\}$ as normally distributed

- The likelihood function of a normal distribution:

$$L(\mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} exp(-\frac{(x_i - \mu)^2}{2\sigma^2})$$

$\hat{\mu}$, $\hat{\sigma}$?

MLE for normal distribution

- Suppose we model the dataset $D = \{x\}$ as normally distributed

- There are two parameters to estimate: ***μ*** and ***σ***
  - If we fix ***σ*** and set θ= ***μ***

    assume σ is known

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

  - If we fix ***μ*** and set θ= ***σ***

    $\mu = \hat{\theta}$

    $$\hat{\theta} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

$$L(\mu,\sigma)$$

fix $\sigma \rightarrow \dfrac{dL(\mu,\sigma)}{d\mu} = 0$

fix $\mu \rightarrow \dfrac{dL(\mu,\sigma)}{d\sigma} = 0$

$$\begin{cases} f_1(\sigma, \hat{\mu}) = 0 \\ f_2(\hat{\sigma}, \mu) = 0 \end{cases}$$

$\hat{\mu}$ $\hat{\sigma}$

In the case of fixing σ. using log Likelihood function $\hat{\mu}$ doesn't depend on σ.

in the case of fixing μ, we assume we know μ using $\hat{\mu}$ as estimation of μ

Drawbacks of MLE

- Maximizing some likelihood or log-likelihood function is mathematically hard
- If there are very few data items, the MLE estimate maybe very unreliable
  - If we observe 3 heads in 10 coin tosses, should we accept that p(heads)= 0.3 ?
  - If we observe 0 heads in 2 coin tosses, should we accept that p(heads)= 0 ?

Confidence intervals for MLE estimates

- An MLE parameter estimate $\hat{\theta}$ depends on the data that was observed

- We can construct a confidence interval for $\hat{\theta}$ using the parametric bootstrap
  - Use the distribution with parameter $\hat{\theta}$ to generate a large number of bootstrap samples
  - From each "synthetic" dataset, re-estimate the parameter using MLE
  - Use the histogram of these re-estimates to construct a confidence interval

Assignments

- Finish Chapter 7 of the textbook
- Next time: Maximum likelihood estimate, Bayesian inference

Additional References

- Robert V. Hogg, Elliot A. Tanis and Dale L. Zimmerman. “Probability and Statistical Inference”
- Morris H. Degroot and Mark J. Schervish "Probability and Statistics”

We are interested in comparing sample means

- Are the average daily body temperature of the two beavers the same?
- We need to model the difference between two sample means

vs.

How do we model the difference between two samples means?

- We know when the sample size ***N*** is large, the sample mean random variable approaches normal *.

- So our problem became **finding the model of the difference between two normally distributed random variables.**

*** Assume the daily temperature at different times are independent.**

Background: sum of independent normals

- We know

$$X_1 \sim normal(\mu_1, \sigma_1^2)$$

$$X_2 \sim normal(\mu_2, \sigma_2^2)$$

$$X_1 + X_2 \sim \quad ?$$

- The sum of X_1 and X_2 is still normal (proof omitted, ref. …)

Background: sum of independent normals

- We know

$$X_1 \sim normal(\mu_1, \sigma_1^2)$$

$$X_2 \sim normal(\mu_2, \sigma_2^2)$$

- **So** $X_1 + X_2 \sim normal(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$
- By the linearity of expected value and the sum rule of variance of the sum of two independent random variables.

Background: sum of independent normals

- We know

$$X_1 \sim normal(\mu_1, \sigma_1^2)$$

$$X_2 \sim normal(\mu_2, \sigma_2^2)$$

- **So** $\boxed{X_1 + X_2 \sim normal(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)}$

- By properties:

$$E[X_1 + X_2] = E[X_1] + E[X_2]$$

$$var[X_1 + X_2] = var[X_1] + var[X_2]$$

Difference of independent normals

- We know

$$X_1 \sim normal(\mu_1, \sigma_1^2)$$

$$X_2 \sim normal(\mu_2, \sigma_2^2)$$

$$X_1 - X_2 \sim \quad ?$$

- The difference of X_1 and X_2 is still normal (proof omitted)

Difference of independent normals

- We know

$$X_1 \sim normal(\mu_1, \sigma_1^2)$$

$$X_2 \sim normal(\mu_2, \sigma_2^2)$$

**

- **So** $\boxed{X_1 - X_2 \sim normal(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)}$

- By the linearity of expected value and the sum rule of variance of the sum of two independent random variables and the scaling property of variance.

Derivation of the mean and variance of difference of independent normals

- Because

- **

$$\boxed{X_1 - X_2 \sim normal(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)}$$

Derivation of the mean and variance of difference of independent normals

- Because $E[X_1 - X_2] = E[X_1] - E[X_2]$

  $= \mu_1 - \mu_2$

Derivation of the mean and variance of difference of independent normals

- Because $E[X_1 - X_2] = E[X_1] - E[X_2]$

$$= \mu_1 - \mu_2$$

$$var[X_1 - X_2] = var[X_1 + (-X_2)]$$

Derivation of the mean and variance of difference of independent normals

- Because $$E[X_1 - X_2] = E[X_1] - E[X_2]$$
  $$= \mu_1 - \mu_2$$

  $$var[X_1 - X_2] = var[X_1 + (-X_2)]$$
  $$= var[X_1] + var[-X_2]$$

Derivation of the mean and variance of difference of independent normals

- Because $E[X_1 - X_2] = E[X_1] - E[X_2]$

$$= \mu_1 - \mu_2$$

$$var[X_1 - X_2] = var[X_1 + (-X_2)]$$

$$= var[X_1] + var[-X_2]$$

$$= var[X_1] + var[X_2]$$

$$var[c \cdot X_2] = c^2 var[X_2]$$

Derivation of the mean and variance of difference of independent normals

- Because

$$E[X_1 - X_2] = E[X_1] - E[X_2]$$
$$= \mu_1 - \mu_2$$

$$var[X_1 - X_2] = var[X_1 + (-X_2)]$$
$$= var[X_1] + var[-X_2]$$
$$= var[X_1] + var[X_2]$$
$$= \sigma_1^2 + \sigma_2^2$$

Derivation of the mean and variance of difference of independent normals

- Because $$E[X_1 - X_2] = E[X_1] - E[X_2]$$
  $$= \mu_1 - \mu_2$$

  $$var[X_1 - X_2] = var[X_1 + (-X_2)]$$
  $$= var[X_1] + var[-X_2]$$
  $$= var[X_1] + var[X_2]$$
  $$= \sigma_1^2 + \sigma_2^2$$

- **

$$\boxed{X_1 - X_2 \sim normal(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)}$$

Now we are ready to check the differences between sample means

- Because sample means are roughly normal when ***N*** is large.

**

$$\boxed{X_1 - X_2 \sim normal(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)}$$

The difference between two sample means

- Suppose we draw samples from two populations $\{x\}$ and $\{y\}$
  - From a sample of size k_x from $\{x\}$, we get sample mean $X^{(k_x)}$
  - From a sample of size k_y from $\{y\}$, we get sample mean $Y^{(k_y)}$

The difference between two sample means

- Define random variable $D = X^{(k_x)} - Y^{(k_y)}$ as the difference between the sample means

- If we hypothesize that popmean($\{x\}$) = popmean($\{y\}$), then

$$E[D] = E[X^{(k_x)}] - E[Y^{(k_y)}] = 0$$

Standard error of the difference between two sample means

- Recall the standard error is roughly the standard deviation of a sample mean

- By the property of variance of the difference between two independent normals

$$var[D] \doteq stderr(\{x\})^2 + stderr(\{y\})^2$$

$$std[D] \doteq \sqrt{stderr(\{x\})^2 + stderr(\{y\})^2} = stderr[D]$$

$$std[D] \doteq \sqrt{\frac{stdunbiased(\{x\})^2}{k_x} + \frac{stdunbiased(\{y\})^2}{k_y}}$$

P-value for testing the equality of two means

- Define the test statistic

$$g = \frac{mean(\{x\}) - mean(\{y\})}{stderr(D)}$$

- If $k_x \geq 30$ and If $k_y \geq 30$

$$\boxed{p = 1 - f = 1 - \frac{1}{\sqrt{2\pi}} \int_{-|g|}^{|g|} exp(-\frac{u^2}{2}) du}$$

P-value: Rejection region- “The extreme fraction”

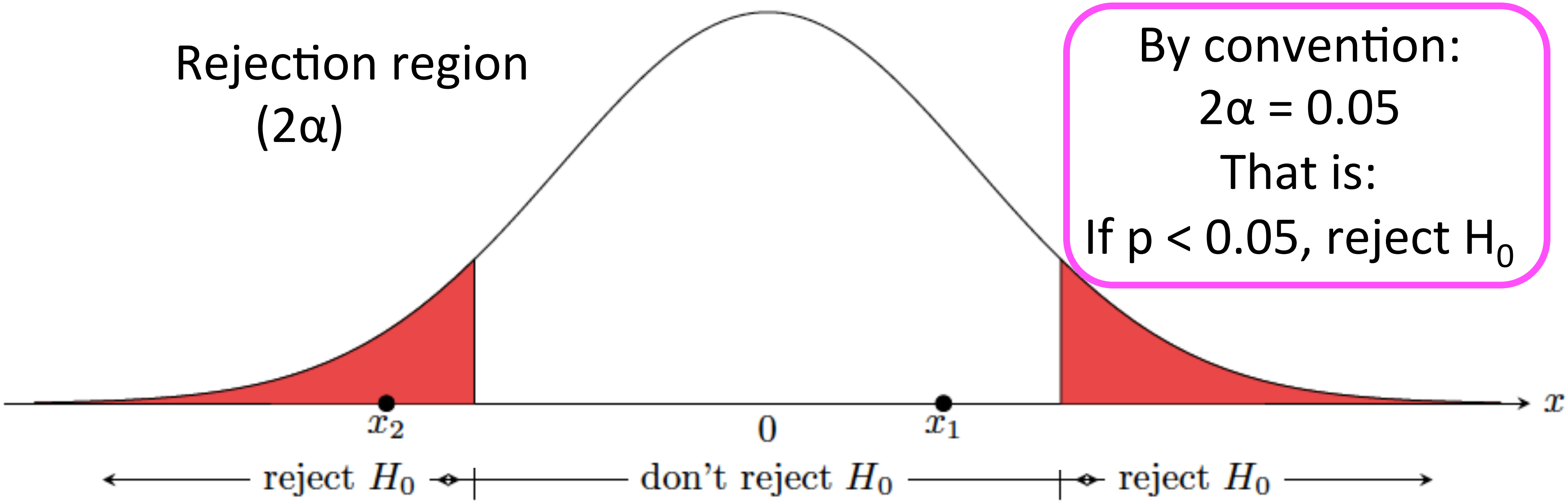
- It is conventional to report the p-value of a hypothesis test

$$p = 1 - f = 1 - \frac{1}{\sqrt{2\pi}} \int_{-|g|}^{|g|} exp(-\frac{u^2}{2})du$$

- Since *N*>30, x should come from a standard normal

Rejection region (2α)

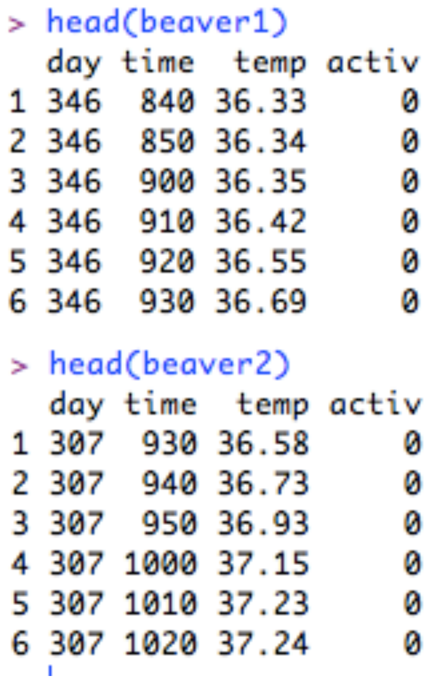
By convention:
2α = 0.05
That is:
If p < 0.05, reject H_0



Comparing the body temperatures of two beavers

- $k_x = 114$ and $k_y = 100$
- Mean($\{x\}$) = 36.86219
- Mean($\{y\}$) = 37.5967
- stderr($\{x\}$) = $\dfrac{stdunbiased(\{x\})}{\sqrt{114}}$
- stderr($\{y\}$) = $\dfrac{stdunbiased(\{y\})}{\sqrt{100}}$
- stderr(D) = $\sqrt{stderr(\{x\})^2 + stderr(\{y\})^2}$

  = 0.04821181



```
> head(beaver1)
  day time  temp activ
1 346  840 36.33     0
2 346  850 36.34     0
3 346  900 36.35     0
4 346  910 36.42     0
5 346  920 36.55     0
6 346  930 36.69     0
```

$\{x\}$

```
> head(beaver2)
  day time  temp activ
1 307  930 36.58     0
2 307  940 36.73     0
3 307  950 36.93     0
4 307 1000 37.15     0
5 307 1010 37.23     0
6 307 1020 37.24     0
```

$\{y\}$

Comparing the body temperatures of two beavers

- Hypothesis H_0: the mean temperatures of the two beavers are the same
- The test statistic g = $\frac{36.86219 - 37.5967}{0.04821181}$ = -15.235

$$p = 1 - f = 1 - \frac{1}{\sqrt{2\pi}} \int_{-15.235}^{15.235} exp(-\frac{u^2}{2}) du$$

$$p \simeq 0$$

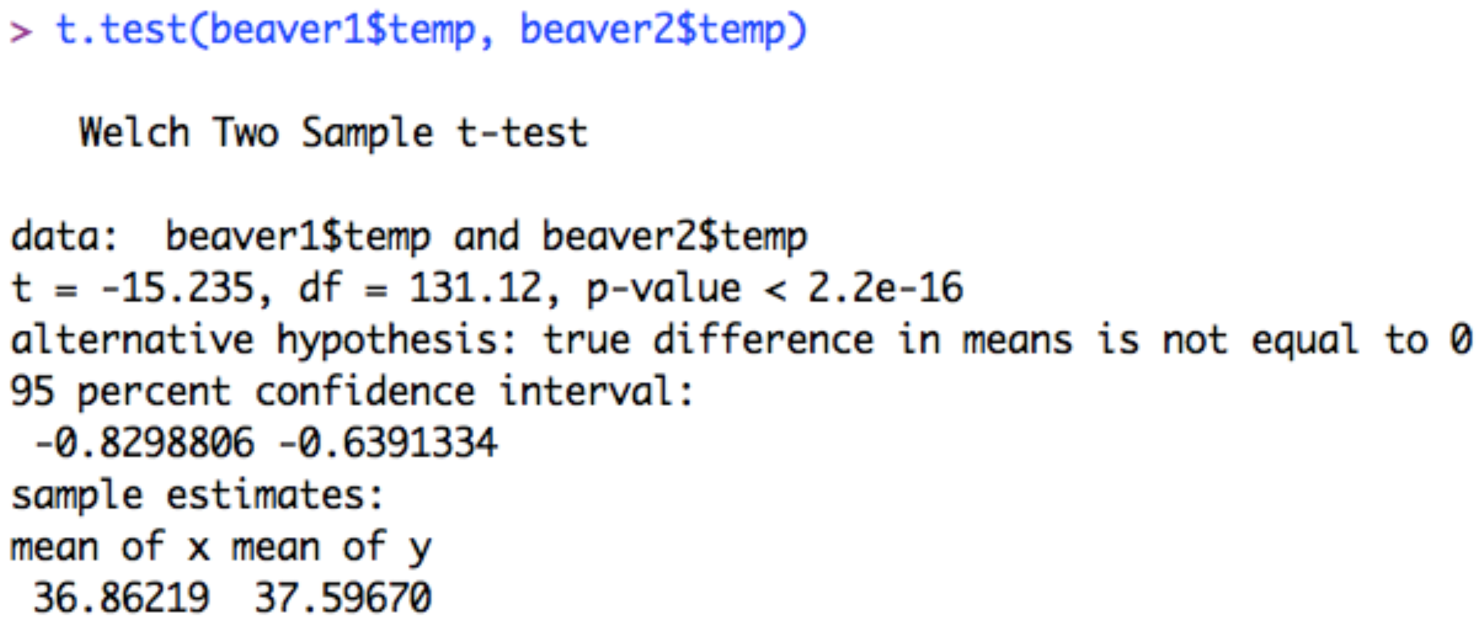
- So we can reject the hypothesis that the mean temperatures are the same

What if N < 30?

- There are general solutions for either N >= 30 or N < 30 if the data sets are random samples from normal distributed data.
  - The difference between sample means can be either modeled as t-distribution with degree ($k_x + k_y - 2$) when their population standard deviations are the same
  - Or the difference between sample means can be approximated with t-distribution with other proper degree of freedom.
  - There are build in t-test procedures in Python, R

Compare the two mean temperatures of two beavers with t.test

- Hypothesis H_0: the mean temperatures of the two beavers are the same



```
> t.test(beaver1$temp, beaver2$temp)

	Welch Two Sample t-test

data:  beaver1$temp and beaver2$temp
t = -15.235, df = 131.12, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.8298806 -0.6391334
sample estimates:
mean of x mean of y 
 36.86219  37.59670 
```

- p < 2.2e-16 , also reject the hypothesis

See you next time

See You!