COMP7/8150: Fundamentals of Data Science

Week 6: Statistical Inference / Estimation

Max H. Garzon

Data Life Cycle

- Problem Definition/goal
  - Identify/specify goals of the data analysis
  - commit to specific deliverables
- Data pre-processing
  - Identify appropriate data
  - Acquire data (gather, lookup, understand)
- Data processing
  - Identify methods (gather, cleanse, store)
  - Carry out the analysis (patterns, trends, predictions?)
- Data post-processing
  - Visualize and present
  - Deploy and evaluate. Iterate, if necessary

Learning Objectives

- To identify the concept of statistical inference and how it is applied
- To apply the Central Limit Theorem (CLT) to derive various sampling distributions
- To identify the concept of a “confidence interval” and how to find them when making inferences with both large (normal distribution) and small sample sizes (Student-t distribution).

A simple statistical model

- Consider the simplest linear statistical model $y = \mu + \varepsilon$
- This model describes a random sample (y_1, y_2, …, y_n) taken from a population with mean μ and standard deviation σ.
- Usually we do not know parameters μ or σ.
- How to provide estimates for the model with
  - A point estimate and its margin of error?
  - An interval estimate with a confidence level?

Another statistical model

- In the first order linear model

  $y = \alpha + \beta x + \varepsilon$

  the coefficients α and β are usually unknown

  Determining them is called "Fitting the model"

- Random error component ε is assumed to follow $N(0, \sigma^2)$

- Key question: how to find the "best estimate" of the parameters α, β, and σ^2 based on observed sample data $(x_1, y_1), (x_2, y_2), \dots (x_n, y_n)$

Model fitting in R

- In R, we use the function lm() to fit most of the linear models in Statistics.
- lm can be used to carry out regression, analysis of variance and analysis of covariance.
- **Usage**
  - lm(formula)
  - lm(formula, data)
  - can get a complete list of argument with help(lm)
- outputs of lm can be further processed by other functions (e.g. summary, plot, anova).

.. Model fitting in R: examples

- Example 1: the simplest statistical model

  $y = \mu + \varepsilon$

  can be fitted with lm(y~1) .

- Example 2: the first order linear model

  $y = \alpha + \beta x + \varepsilon$

  can be fitted by lm(y~x)

- Example 3: the first order linear model with zero intercept $y = \beta x + \varepsilon$

  can be fitted by lm(y~x-1) .

Types of Statistical Inference

- **Estimation:**
  - Estimating or predicting the value of the parameter.
  - Provide the accuracy of the estimate

- **Hypothesis Testing:**
  - Making a “sound decision” to decide: “Did the sample come from a specific type of population?”

What is estimation?

http://www.youtube.com/watch?v=mD56-raCdGg&feature=related

- Consider a statistical model to describe a random sample of size n taken from a population distribution with certain parameters of interest (e.g. mean μ and standard deviation σ .)
- Usually, we do not know these parameters
- The problem of estimation is to provide approximations of the actual values of the unknown parameters, such as μ and σ .
- Can be done in two ways:
  - a point estimate and its margin of error; or/and as
  - an interval estimate with a confidence level.

Confidence Interval Estimation

- Create an interval (a, b) so that you are fairly sure (with certain confidence level) that the actual parameter lies between these two values.
- With confidence coefficient 1- α that the interval (a,b) will cover the true parameter.
- It is common to use 1- α = .95 (α =0.05) and assume that the estimator has a normal distribution.

The Margin of Error

- Margin of error
  The maximum error of estimation, calculated as 1.96 SE, where SE is the standard error (standard deviation) of the estimator used for the given parameter.

- Why?
  For ***unbiased*** estimators with (approximate) normal sampling distributions, 95% of all point estimates will lie within 1.96 SE of the parameter of interest.

Computation of SE

- The computation of SE for an estimator can be hard in some cases but for most common cases, formulas are available.

- Most common cases of estimation
  1. Mean μ of a population
  2. proportion (p) of a binomial distribution B(n,p).
  3. difference of means of two populations $\mu_1 - \mu_2$.
  4. difference of two proportions in binomial distributions $B(n_1,p_1)$ and $B(n_2,p_2)$ $\mathbf{p_1 - p_2}$

Case 1: Estimating the Mean

http://www.youtube.com/watch?v=x6OsGXwi1hU
http://www.youtube.com/watch?v=zBASlmIfR9s&feature=fvwrel

Population	mean μ and s.d. σ
Parameter of interest	μ
Sample	random sample of size n
Sample statistics	sample mean $\bar{x}$, sample s.d. s
Point estimator of μ	$\bar{x}$
Standard Error of $(\bar{x})$	$\mathrm{SE} = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}}$
Margin of error	$\pm 1.96\mathrm{SE} \approx \pm 1.96\frac{s}{\sqrt{n}}$
$100(1-\alpha)\%$ C. I.	$\bar{x} \pm z_{\alpha/2}\mathrm{SE} \approx \bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}$

.. Case 1: Estimating Mean

- Suppose that we want to estimate the actual amount of soda in a 12 oz. can. We randomly samples 200 cans of soda and find average fill is 11.9oz with a standard deviation of 0.2oz.
- Find a 95% confidence interval for the true level of soda in the can.
- Using R

```
#1.96 = qnorm(0.975)=qnorm(1-0.05/2)
> 11.9 +c(-1,1)*qnorm(0.975)*0.2/200^0.5
[1] 11.87228 11.92772
```

.. Estimating Mean in R

- Suppose that the actual amount of soda is indeed 12 oz. in a 12 oz. can with a standard deviation of 0.2oz.
- Write R code to simulate random sampling of 200 cans of soda and find sample average fill.
- Run the code1000 times (in a for loop) and for each sample.
- Can you find the percentage of time that the computed 95% confidence interval actually covers the true level? Does 950 times have anything to do with it? Why so?

Case 2: Estimating Binomial Proportion

Population	$X \sim B(n, \pi)$
Parameter of interest	π
Sample statistics	x
Point estimator of π	$\hat{\pi} = x/n$
Standard Error of $(\hat{\pi})$	$\mathrm{SE} = \sqrt{\frac{\pi(1-\pi)}{n}} \approx \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$
Margin of error	$\pm 1.96\mathrm{SE} \approx \pm 1.96\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$
$100(1-\alpha)\%$ C. I.	$\hat{\pi} \pm z_{\alpha/2}\mathrm{SE} \approx \hat{\pi} \pm z_{\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$

.. Case 2: Estimating Binomial Proportion

- Suppose that we want to estimate the proportion of soda cans that are underfilled (less than 12 oz).
- We randomly samples 200 cans of soda and finds 10 underfilled cans.
- Using R, find a 95% confidence interval for the true percentage of underfilled cans:

```
> p <- 10/200
> p + c(-1,1)*qnorm(0.975)*(p*(1-
  p)/200)^0.5
[1] 0.01979493 0.08020507
```

Case 3: Estimating difference $\mu_1 - \mu_2$

- For two quantitative populations with unknown means (μ_1, μ_2) and standard deviations (σ_1, σ_2) .
- We take a random sample of sizes n_1, n_2 from the two populations, and compute their sample means $\bar{x}_1, \bar{x}_2$ and sample standard deviations (s_1, s_2).
- We are interested in the sampling distribution of $\bar{x}_1 - \bar{x}_2$

.. Case 3: Estimating difference $\mu_1 - \mu_2$

Populations	pop 1: $(\mu_1$ and $\sigma_1)$, pop 2: $(\mu_2$ and $\sigma_2)$
Parameter of interest	$\mu_1 - \mu_2$
Samples	sample of size n_1 from pop 1, size n_2 from pop 2
Sample statistics	sample mean $\bar{x}_1$, $\bar{x}_2$ and sample s.d. s_1 and s_2
Point estimator of $\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$
Standard Error of $(\bar{x}_1 - \bar{x}_2)$	$\text{SE} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \approx \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
Margin of error	$\pm 1.96\text{SE} \approx \pm 1.96\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
$100(1-\alpha)\%$ C. I.	$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\text{SE} \approx (\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

.. Case 3: Estimating difference $\mu_1 - \mu_2$

- Two types of bottling machines (A and B) are used to fill soda cans and the sample data are shown in the table below.
- Find a 95% confidence interval for the difference of level filled between two machines.
- R-code:

```
> (12.1-11.9) + c(-1,1)*qnorm(0.975)*(0.25^2/100+0.15^2/100)^0.5
[1] 0.1428577 0.2571423
```

Level filled in cans	A	B
Sample size	100	100
Sample mean	12.1	11.9
Sample Std Dev	0.25	0.15

Case 4: Estimating the difference between two proportions

- Examples to compare the proportion of "successes" in two binomial populations.
  1. The germination rates of untreated seeds and seeds treated with a fungicide.
  2. The proportion of male and female voters who favor a particular candidate for governor.

.. Case 4: Estimating the difference between two proportions

Populations	$X_1 \sim B(n_1, \pi_1)$ and $X_2 \sim B(n_2, \pi_2)$
Parameter of interest	$\pi_1 - \pi_2$
Sample statistics	x_1 from $B(n_1, \pi_1)$ and x_2 from $B(n_2, \pi_2)$
Point estimator of $\pi_1 - \pi_2$	$\hat{\pi}_1 - \hat{\pi}_2 = x_1/n_1 - x_2/n_2$
Standard Error of $(\hat{\pi}_1 - \hat{\pi}_2)$	$\text{SE} = \sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}} \approx \sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}$
Margin of error	$\pm 1.96\text{SE} \approx \pm 1.96\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}$
$100(1-\alpha)\%$ C. I.	$\hat{\pi}_1 - \hat{\pi}_2 \pm z_{\alpha/2}\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}$

.. Case 4: Estimating the difference between two proportions

- Two types of bottling machines A, B are used to fill the soda cans; a sample of 100 cans taken for each machine yields a number defective cans of 4, 7.
- Find a 95% confidence interval for the difference of percentage of underfilled between A and B
- Using R:

```
> p1 <- 4/100; p2 <- 7/100;
> (p1-p2)+ c(-1,1)*qnorm(0.975)*(p1*(1-
p1)/100+p2*(1-p2)/100)^0.5
[1] -0.09305482  0.03305482
```

General Case:
Sampling distributions in general

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

When the sample size is large, then the sampling distribution of z is N(0,1).

$$z = \frac{\hat{\theta} - \theta}{SE(\hat{\theta})}$$

For a general parameter and its estimator, we have a similar result on the sampling distribution of z, N(0,1).
The key problem is to find its SE.

Constructing Confidence Interval

http://www.youtube.com/watch?v=bq9XhIM0gAQ&feature=related

- When the sample size is large, the sampling distribution of z is N(0,1). Let
  $z_{\alpha/2}$ = percentile of the N(0,1) distribution.

$$z = \frac{\hat{\theta} - \theta}{SE(\hat{\theta})}$$

- We can construct a 100(1-α)% confidence interval for θ as

$$\hat{\theta} \pm z_{\alpha/2} SE(\hat{\theta})$$

Summary of four common cases for large sample sizes

Parameter	Estimator	Margin of Error	$100(1-\alpha)\%$ Confidence Interval
μ	$\bar{x}$	$\pm 1.96 \frac{s}{\sqrt{n}}$	$\bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}}$
p	$\hat{p} = x/n$	$\pm 1.96 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
$\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$	$\pm 1.96 \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$	$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
$p_1 - p_2$	$\hat{p}_1 - \hat{p}_2$	$\pm 1.96 \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$	$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

Statistical Inference with small samples

- So far, we rely on CLT when the sample sizes are large so can assume normality.
- What if the sample sizes are not large ?
- For the case of binomial distribution, exact small sample inference is available.

- For a quantitative population with unknown mean μ **and standard deviation** σ, we need small sample theory.

Small Sample Inference

- Consider a quantitative population with unknown mean μ and standard deviation σ.
- We take a random sample of size n and compute the sample mean $\bar{x}$ and sample standard deviation s.
- We need the normality assumption on the population to proceed as before

Types of small sample inference

- When the sample size is small, the estimation and testing procedures obtained from CLT are not appropriate.

- Common small sample inferences for
  - ✓ μ, the mean of a normal population
  - ✓ $\mu_1 - \mu_2$, the difference between two population means
  - ✓ $\mu_1 - \mu_2$, the difference between one paired population means

Sampling Distribution of z and t Stats

- Assuming a normal population, the sample mean $\bar{x}$ has a normal distribution for any sample size *n*, and the z-statistic has a standard normal distribution.
- But if σ is unknown (so we must use *s* to estimate it), the t statistic may not be normal.

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

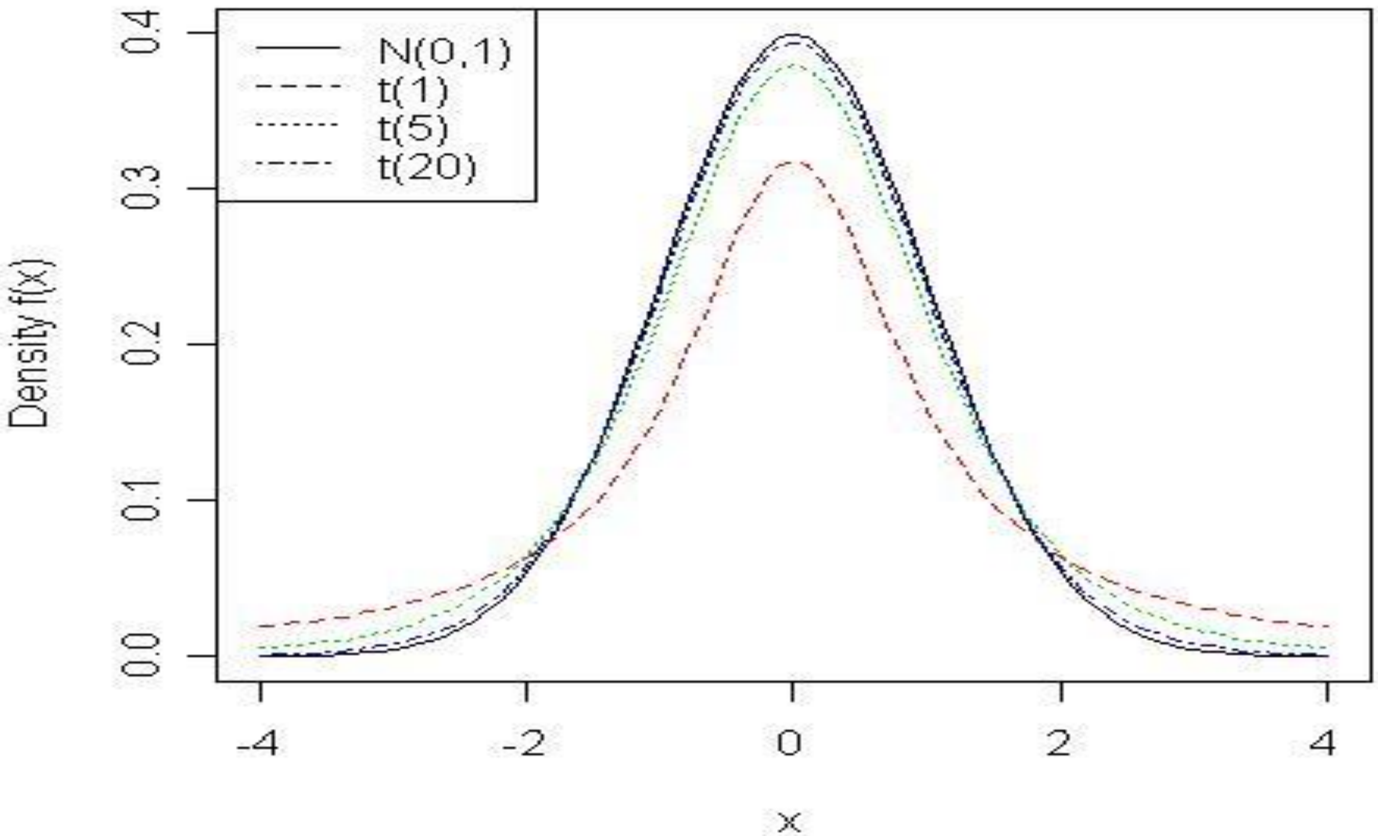
.. Student's t Distribution

http://www.youtube.com/watch?v=NACUg0PdjIc

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

- The sampling distribution of t is known as the Student's t distribution, with *n*-1 degrees of freedom.
- We can create estimation or testing procedures for the population mean μ.

.. Student's t Distribution: Sample Plots



.. Student's t distribution

t distribution

$X \sim t(v)$ with p.d.f.

$$p(x) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\Gamma\left(\frac{v}{2}\right)\sqrt{v\pi}}\left(1 + \frac{x^2}{v}\right)^{-\frac{v+1}{2}}, \quad -\infty < x < \infty.$$

1. In R: df=v, ncp=0 (default) distribution function=t.
2. $E(X) = 0$ (when $v > 1$).
3. $Var(X) = \frac{v}{v-2}$ (when $v > 2$).

Computing probability in R

- Recall that, in general, for a distribution named xxx, you can use following R code
  - dxxx = probability distribution function, $P(X=x)$.
  - **p**xxx = cumulative probability distribution, $P(X \le x)$.
  - **q**xxx = percentile of probability distribution, finding the smallest x value that $P(X \le x) \ge q$.
  - **r**xxx = generate a random number from xxx distribution.
- For N(0,1), xxx=norm; for t(df), xxx=t.

Computing Student's t in R

- Syntax:
  - `dt(x, df, ncp, log = FALSE)`
  - `pt(q, df, ncp, lower.tail = TRUE, log.p = FALSE)`
  - `qt(p, df, ncp, lower.tail = TRUE, log.p = FALSE)`
  - `rt(n, df, ncp)`
  - Note: default values can be (and usually are) omitted. The third parameter is ncp (non-centrality parameter) is generally omitted

R code

```
x <- c(-40:40)*0.1
y1 <- dnorm(x,0,1)
y2 <- dt(x,1)
y3 <- dt(x,5)
y4 <- dt(x,20)
y <- cbind(y1,y2, y3, y4)
matplot(x, y, type="l", ylab="Density f(x)")
legend("topleft",
c("N(0,1)","t(1)","t(5)","t(20)"),
lty=c(1,2,3,4))
#as df gets larger and larger, it is closer to
N(0,1).
```

Normal vs. t Distribution

- Both N(0,1) and t-distribution are symmetric around 0. The shapes of their distributions are similar.
- In the next two slides, we will show that the difference between the distributions of N(0,1) and Student's t with df=v.
- The tail of Student's t is heavier than that of N(0,1). Hence, the upper percentile is larger for the t distribution.
- The difference is getting smaller as df=v gets larger. This is consistent with the previous plot.

.. Normal vs. t Distribution-Example 1

- Using R, find 95% percentile of N(0,1) and various t-distributions:
  - ```
    > options(digits=5)   # set digits for display
    > qnorm(0.95)
    [1] 1.6449
    ```
  - ```
    > qt(0.95, 1)
    [1] 6.3138
    ```
  - ```
    > qt(0.95, 5)
    [1] 2.0150
    ```
  - ```
    > qt(0.95, 20)
    [1] 1.7247
    ```

.. Normal vs. t Distribution-Example 2

- Using R to find Pr(X ≤ 1) when X follows N(0,1) and various t-distributions:
  - ```
    > pnorm(1)
    [1] 0.84134
    ```
  - ```
    > pt(1, 1)
    [1] 0.75
    ```
  - ```
    > pt(1, 5)
    [1] 0.81839
    ```
  - ```
    > pt(1, 20)
    [1] 0.83537
    ```
  - ```
    > pt(1,50)
    [1] 0.8389372
    ```
    ```

Table 4.1 and Fig 4.2: comparing the critical values of N(0,1) and t(df)

- >#We can reproduce the output from Table 4.1 (page 107)
- >#note that t with df=30 has much closer values to N(0,1)
- > p <- c(0.841, 0.975, 0.995, 0.9995)
- > qnorm(p)

  [1] 0.9985763 1.9599640 2.5758293 3.2905267
- > qt(p, 2)

  [1] 1.318781 4.302653 9.924843 31.599055
- > qt(p, 8)

  [1] 1.064908 2.306004 3.355387 5.041305
- > qt(p, 30)

  [1] 1.015474 2.042272 2.749996 3.645959

Constructing Confidence Intervals with small samples

$$t = \frac{\hat{\theta} - \theta}{SE(\hat{\theta})}$$

When sample size is small, the sampling distribution of t is Student's t distribution with degrees of freedom v. Define

$t_{1-\alpha/2,v}$ = upper percentile of t(v) distribution.

We can construct a 100(1-α)% confidence interval for θ as

$$\hat{\theta} \pm t_{1-\alpha/2,v} SE(\hat{\theta})$$

Critical values of normal and t distributions

- To construct a confidence interval, we need to find the critical value like
  - $z_{\alpha/2}$ = upper $\alpha/2$ percentile of N(0,1) distribution (when sample size is large) .
  - $t_{1-\alpha/2,v}$= upper $\alpha/2$ percentile of t_v distribution (when the sample size is small) where v is the degrees of freedom.

$$\hat{\theta} \pm z_{\alpha/2} SE(\hat{\theta})$$

$$\hat{\theta} \pm t_{1-\alpha/2,v} SE(\hat{\theta})$$

.. Critical values of normal and t distributions

- Since normal and t distributions are symmetric about 0, we have $z_{1-\alpha/2} = -z_{\alpha/2}$ and $t_{1-\alpha/2,v} = -t_{\alpha/2,v}$
- It is common to use $z_{\alpha/2}$ as the upper percentile.
- However, the notation used in our textbook for upper percentile of t is $t_{1-\alpha/2,v}$, not $t_{\alpha/2,v}$

```
> qt(0.1, 8)
[1] -1.396815
> qt(0.9, 8)
[1] 1.396815
```

- (confirm this fact using R)
- `> qnorm(0.1)`

  `[1] -1.281552`
- `> qnorm(0.9)`

Case 1S: Small Sample Inference for a Population Mean μ

Population	Normal with mean μ and s.d. σ
Parameter of interest	μ
Sample	random sample of size n
Sample statistics	sample mean $\bar{x}$, sample s.d. s
Point estimator of μ	$\bar{x}$
Standard Error of $(\bar{x})$	$\mathrm{SE} = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}}$
$100(1-\alpha)\%$ C. I.	$\bar{x} \pm t\mathrm{SE} \approx \bar{x} \pm t\frac{s}{\sqrt{n}}, \quad \mathrm{t}=\mathrm{t}_{1-\alpha/2,n-1}$

.. Case 1S: Small Sample Inference for a Population Mean μ: example

- Suppose that we want to estimate the average amount of soda in a 12 oz. can. We randomly sample 6 cans of soda and find sample average fill is 11.9oz with a standard deviation of 0.2oz.
- Find a 95% confidence interval for the true level of soda in the can.
- R-code:

```
# small sample  size with df=6-1=5
 > 11.9 + c(-1,1)*qt(0.975,5)*0.2/6^0.5
 [1] 11.69011 12.10989
```

Case 2S: Estimating the Difference between Two Means

Populations	pop 1: $N(\mu_1, \sigma^2)$, pop 2: $N(\mu_2, \sigma^2)$
Parameter of interest	$\mu_1 - \mu_2$
Samples	sample of size n_1 from pop 1, size n_2 from pop 2
Sample statistics	sample mean $\bar{x}_1$, $\bar{x}_2$ and sample s.d. s_1 and s_2
Point estimator of $\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$
Polled variance	$s_p^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$
Standard Error of $(\bar{x}_1 - \bar{x}_2)$	$\mathrm{SE} = \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \approx s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
$100(1-\alpha)\%$ C. I.	$(\bar{x}_1 - \bar{x}_2) \pm t s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \quad t = t_{1-\alpha/2, n_1+n_2-2}$

.. Case 2S: Estimating the difference $\mu_1 - \mu_2$

- Two types of bottling machines (A and B) are used to fill soda cans and the sample data are shown in the table below.
- Find a 95% confidence interval for the difference of level filled between two machines.
- R-code:

```
> n1 <- 6; n2 <- 12; s1 <- 0.25; s2 <- 0.15;
> sp2 <- ((n1-1)*s1^2+(n2-1)*s2^2)/(n1+n2-2);
> sp <- sp2^0.5; df <- n1+n2-2;
> (12.1-11.9)+c(-1,1)*qt(0.975,df)*sp*(1/n1+1/n2)^0.5
[1] 0.001701017 0.398298983
```

Level filled in cans	A	B
Sample size	6	12
Sample mean	12.1	11.9
Sample Std Dev	0.25	0.15

Case 3S: Estimating the mean of a Paired Difference

- For the previous Case 2s, we assumed that the data were sampled independently from two populations: sample size n_1 from population 1 and sample size n_2 from population 2. That is called a two-sample design
  - For certain experiments, this independence assumption is clearly violated.
- To minimize the variation in an experiment, we can consider a paired-sample design where two different measurements are taken on same unit.

..Case 3S: Estimating the mean of a Paired Difference

- To measure the effectiveness of certain blood pressure drug, we measure the individual's blood pressure before and after taking the drug. Clearly, the two measurements (before and after) on the same person are not independent.
- To measure the level of soda can filling, we can use two measuring methods: (1) estimate it based on its weight without opening; or (2) actually measure it. Again, two measurements on the same can are not independent.

..Case 3S: Estimating the mean of a Paired Difference

- In previous examples, the design used is a paired design where two different (but not independent) measures were taken on the sample unit.
  - Usually, we can have a more precise estimate on the difference.
- For the sample problem, we can also consider a two-sample design, but that is usually less precise.
  - Example: we take a sample from a population without taking the drug and then take another sample from a population with drug treatment.

Case 3S: Estimating the mean of Paired Difference

Population	(X_1, X_2) pair with $D = X_1 - X_2 \sim N(\mu_D, \sigma_D^2)$
Parameter of interest	$\mu_D = \mu_1 - \mu_2$
Sample	n pairs of $(X_{1,i}, X_{2,i})$, $D_i = X_{1,i} - X_{2,i}$
Sample statistics	sample mean $\bar{d} = \bar{x}_1 - \bar{x}_2$, sample s.d. s_d
Point estimator of μ_D	$\bar{d} = \bar{x}_1 - \bar{x}_2$
Standard Error of $(\bar{d})$	$\mathrm{SE} = \frac{\sigma_D}{\sqrt{n}} \approx \frac{s_d}{\sqrt{n}}$
$100(1-\alpha)\%$ C. I.	$\bar{d} \pm t\mathrm{SE} \approx \bar{d} \pm t\frac{s_d}{\sqrt{n}}, \quad t = t_{1-\alpha/2, n-1}$

Constructing confidence intervals in R

- There is a general function in R, called confint(), that can be useful to construct a confidence interval for parameters in a statistical model.
- This function confint() is most useful for more complicated models (to be discussed later), not that useful for simple models discussed in this module.
- You can find how to use this function confint() in R by typing
- ?confint gets your help from R with this command.

Further Topics

- Some of the topics listed will be discussed in the next module.
- Hypothesis testing
- p-value and hypothesis testing
- Contingency tables
- One-way ANOVA table
- Response curves
- Re-sampling methods

Questions?