R inference examples

CLT, CIs, and Tests

Stat 251 Spring 2019

General note

All you really have to do is copy, paste, and modify the code that is found in the following examples. The modifications will mostly be just the summarized numbers or your data (recycle). Your mantra should be "copy, paste, modify, run!"

`R` tools

(1) **Calculate the probability of a z-score**: `pnorm(x,mu,se)` or `pnorm(z)`
- x: x-bar ($\bar{x}$) or p-hat ($\hat{p}$), the argument from the problem
- mu: μ or p, the mean (or proportion) of the distribution
- se: $se_{mean} = stdev/\sqrt{n}$ or $se_{proportion} = \sqrt{pq/n}$
- z: only input z if you calculate a z-score, like in the CLT examples; the examples shown will use `pnorm()` both ways

The `pnorm()` (and `pt()`, `qnorm()`, and `qt()`) function works just like the z table in your table packet; that is, it calculates the area to the *left* of the z-score by default. If you want area to the right, you will use `1-pnorm()`, and area between two values would use `pnorm()-pnorm()`.

(2) **Calculate the probability of a t-score**: `pt(x,mu,se,df)` or `pt(t,df)`
- x: x-bar ($\bar{x}$), the argument from the problem
- mu: μ, the mean of the distribution
- se: $se_{mean} = s/\sqrt{n}$
- df: $df = n - 1$ for 1-sample analyses
- t: only input t if you calculate a t-score

(3) **Find a z-score with given probability (for CIs)**: `qnorm(perc)`
- perc: the percent (as in 0.9 for the 90th percentile) or $1 - \alpha/2$

(4) **Find a t-score with given probability (for CIs)**: `qt(perc,df)`
- perc: the percent (as in 0.9 for the 90th percentile) or $1 - \alpha/2$
- df: $df = n - 1$ for 1-sample analyses

(5) **CI and hypothesis test using t (only works if you have the data values, not summarized statistics)**: `t.test(data,mu,conf.level,alternative)`
- x: dataset (variable name)
- mu: μ_0, the mean from the null hypothesis (H_0)
- conf.level: the confidence level, $1 - \alpha = CL$ (since $1 - CL = \alpha$)
- alternative: by default it is 2-tailed (which is needed for a CI); for $H_a :>$ `alternative='g'` and for $H_a :<$ `alterative='less'`

Probabilities using CLT

CLT in R: proportion example

Based on past experience, a bank believes that 7% of the people who receive loans will not make payments on time. The bank has recently approved 200 loans. (1) What are the mean (mean proportion) and standard error of the proportion of the 200 clients who may not make timely payments? (2) What is the probability that over 10% of these clients will not make timely payments? (3) What is the probability that between 5% and 8% of these clients will not make timely payments?

(1) mean and standard error: mean: $p = 0.07$ and $se = \sqrt{\frac{pq}{n}}$

```
# mean is p
p=0.07; q=1-p
n=200
# se=sqrt(p*q/n)
se=sqrt(p*q/n)
rbind(p,q,n,se)
```

```
            [,1]
p     0.07000000
q     0.93000000
n   200.00000000
se    0.01804162
```

$$\therefore \hat{p} \sim N(0.07, 0.0180416)$$

(2) $P(\hat{p} > 0.10)$: the `pnorm(phat,p,se)` function allows input of mean and se

```
# P(phat>0.1)
# probability of area to the right: P(Z>zcalc)=1-P(Z<zcalc)
# pnorm(phat,p,se)
1-pnorm(0.1,p,se)
```

```
[1] 0.04817404
```

```
# or just input all the numbers as:
1-pnorm(0.1,0.07,0.018)
```

```
[1] 0.04779035
```

(3) $P(0.05 < \hat{p} < 0.08)$

```
# take difference of 2 pnorms (one for 5% and one for 8%)
pnorm(0.08,p,se)-pnorm(0.05,p,se)
```

```
[1] 0.5764917
```

CLT in R: mean example

Planes cannot fly well (or as safely) if the payload is too great. Suppose that an airline runs a commuter flight from New York to Boston and holds up to 40 passengers. The airline knows that the average weight of passenger plus luggage for typical customers on this type of flight is approximately normal with a mean of 225 pounds and standard deviation of 35 pounds (per person). Suppose that a random sample of 40 customers was taken. (1) What are the mean and standard error of the mean weight per passenger plus luggage? (2)

What is the probability that the mean weight will be less than 220 pounds? (3) What is the probability that the mean weight per passenger is between 215 and 220 pounds?

(1) mean and standard error: mean: μ and $se = \frac{\sigma}{\sqrt{n}}$

```
mu=225; sigma=35; n=40
# se=sigma/sqrt(n)
se=sigma/sqrt(n)
rbind(mu,se)
```

```
         [,1]
mu 225.000000
se   5.533986
```

$$\therefore \; \overline{X} \sim N(225, 5.53)$$

(2) $P(\overline{X} < 220)$: the `pnorm(xbar,mu,se)` function will allow input of mean and se

```
pnorm(220,mu,se)
```

```
[1] 0.1831282
```

(3) $P(215 < \overline{X} < 220)$

```
# take difference of 2 pnorms (one for 220 and one for 215)
pnorm(220,mu,se)-pnorm(215,mu,se)
```

```
[1] 0.1477483
```

CLT in R: total example

Planes cannot fly well (or as safely) if the payload is too great. Suppose that an airline runs a commuter flight from New York to Boston and holds up to 40 passengers. The airline knows that the average weight of passenger plus luggage for typical customers on this type of flight is approximately normal with a mean of 225 pounds and standard deviation of 35 pounds (per person). Suppose that a random sample of 40 customers was taken. (1) What are the *total* (mean total) and standard error of the total weight of (40) passengers plus luggage? (2) If the total weight of passengers and luggage cannot exceed 9500 pounds, what is the probability that a sold-out flight (of 40 passengers) will exceed the total weight limit?

(1) mean and standard error: mean: $\tau = n\mu$ and $se = \sqrt{n}(\sigma)$

```
mu=225; n=40; tau=n*mu
sigma=35; se=sigma*sqrt(n)
rbind(tau,se)
```

```
         [,1]
tau 9000.0000
se   221.3594
```

$$\therefore \hat{\tau} \sim N(9000, 221.36)$$

(2) $P(\hat{\tau} > 9500)$: the `pnorm(tauhat,tau,se)` function will allow input of mean and se

```
1-pnorm(9500,tau,se)
```

```
[1] 0.01194886
```

One-sample CIs

CI example for mean with known sigma (σ)

A survey to find the average starting salary for computer science majors was given by the National Association of Colleges and Employers. It is known from previous studies that the standard deviation of starting salaries is \$4300. A random sample of 35 companies was taken and the mean starting salary was \$60,038. Estimate the true mean starting salary of CS majors with 95% confidence.

The mean $\mu = 60038$ and $se = \frac{\sigma}{\sqrt{n}}$

```
xbar=60038; sigma=4300; n=35
se=sigma/sqrt(n); se
```

```
[1] 726.8327
```

```
alpha=0.05
# the negative sign in the next calcualtion is because if you look up z with
# probability=alpha/2, (eg: 0.05/2=0.025) the z-score would technically be negative
# using the negative sign will make the z-score positive
zstar=-qnorm(alpha/2); zstar
```

```
[1] 1.959964
```

```
bound=zstar*se; bound
```

```
[1] 1424.566
```

```
lower=xbar-bound; upper=xbar+bound
rbind(lower,upper)
```

```
          [,1]
lower 58613.43
upper 61462.57
```

```
# try with a negative z-score if you want to see the effect :-)
```

CI example for proportion (z)

A random sample of 200 students from a college are asked if they regularly eat breakfast. Eighty-four students responded that they did eat breakfast regularly. Estimate the true proportion of college students that eat breakfast with 90% confidence.

Calculate $\hat{p}$ with the successes (84) divided by the total (200)

```
n=200; x=84
phat=x/n; phat
```

```
[1] 0.42
```

```
qhat=1-phat; qhat
```

```
[1] 0.58
```

```
alpha=0.1
zstar=-qnorm(alpha/2); zstar
```

```
[1] 1.644854
```

```
se=sqrt(phat*qhat/n); se
```

```
[1] 0.03489986
```

```r
bound=zstar*se; bound
```

```
[1] 0.05740516
```

```r
lower=phat-bound; upper=phat+bound
rbind(lower,upper)
```

```
           [,1]
lower 0.3625948
upper 0.4774052
```

CI example for mean with data and unknown σ (compute s)

Data representing possums in Australia and New Guinea were measured. Measurements included sex, age, head length, skull width, total length and tail length. Of interest is the total length, measured in *cm*. Estimate the true mean total length of possums with 99% confidence.

```r
totlength
```

```
 [1] 91.5 85.0 84.5 82.0 82.5 84.0 89.5 89.0 89.0 86.0 81.0 95.5 87.0 93.5
[15] 87.5 84.0 84.0 89.5 85.0 92.0 85.0 81.0 77.0 84.5 85.0
```

```r
# length() finds the number of observations in the variable, aka the sample size
n=length(totlength)
t.test(totlength,conf.level=0.99)
```

```

    One Sample t-test

data:  totlength
t = 100.93, df = 24, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
99 percent confidence interval:
 83.79182 88.56818
sample estimates:
mean of x
    86.18
```

We are 99% confident the true mean total length of possums is between 83.791822 and 88.568178 *cm*.

If no data is given and only summary statistics (mean, standard deviation, and sample size), then you have to calculate it the same way as we did the CI with z in a previous example here.

99% CI for mean with no data (summary statistics only) and unknown sigma (using s in summary statistics).

Same example with the possums:

You will have to assign your xbar, s, and n (example: `xbar=54; s=9; n=20`). The following example's mean, standard deviation, and sample size are a part of a random sample so I cannot assign them here in this example since they will change everytime I run this code (i.e)

```
xbar; s; n
```

```
[1] 86.18
```

```
[1] 4.26927
```

```
[1] 25
```

```
rbind(xbar,s,n)
```

```
         [,1]
xbar 86.18000
s     4.26927
n    25.00000
```

```
alpha=0.01; se=s/sqrt(n); se
```

```
[1] 0.853854
```

```
tstar=-qt(alpha/2,df=n-1); tstar
```

```
[1] 2.79694
```

```
bound=tstar*se; bound
```

```
[1] 2.388178
```

```
lower=xbar-bound; upper=xbar+bound
rbind(lower,upper)
```

```
          [,1]
lower 83.79182
upper 88.56818
```

We are 99% confident the true mean total length of possums is between 83.791822 and 88.568178 cm.[1]

[1] Results of the `t.test()` CI and the one done with summary statistics in this example vary because I took a sample of the original dataset for this example, which means every time I run this document, the sample changes because it is a random sample using the command `sample()`. In your assignment, this will not happen because you are not going to use the `sample()` command for anything.

Hypotheses tests

Hypothesis test for mean with known σ

A manufacturer of sprinkler systems used for fire protection in office buildings claims that the true average system-activation temperature is 130° F. It is known from previous studies that the temperatures are normally distributed with standard deviation 1.5° F. A sample of $n = 9$ systems, when tested yields a sample average activation temperature of 131.08° F. Is there sufficient evidence that the true mean activation temperature is more than the manufacturer claim?

$$H_0 : \mu = 130 \ vs. \ H_a : \mu > 130$$

```
mu0=130; xbar=131.08; sigma=1.5; n=9
se=sigma/sqrt(n); se
```

```
[1] 0.5
```

```
zcalc=(xbar-mu0)/se; zcalc
```

```
[1] 2.16
```

```
# Ha > so need to subtract probability from 1
pvalue=1-pnorm(zcalc); pvalue
```

```
[1] 0.01538633
```

We will reject H_0 if $pvalue \leq \alpha(0.05)$. Since $pvalue = 0.0154 \leq \alpha(0.05)$,we reject H_0 and the average activation temperature is more than the claim.

Hypothesis test for mean with unknown σ (using s)

Product in cans are marked as 38 oz. is there sufficient evidence that the can weights are different than what the label states?

$$H_0 : \mu = 38 \ vs. \ H_a : \mu \neq 38$$

```
cans=c(34.6,39.65,34.75,40,39.5,38.9,34.25,36.8,39,37.2)
mu0=38; n=length(cans)
# t.test(data,df,mu=mu0)
# careful to make sure you have correct syntax
t.test(cans,mu=mu0)
```

```

    One Sample t-test

data:  cans
t = -0.74683, df = 9, p-value = 0.4742
alternative hypothesis: true mean is not equal to 38
95 percent confidence interval:
 35.84448 39.08552
sample estimates:
mean of x
   37.465
```

We will reject H_0 if $pvalue \leq \alpha(0.05)$. Since $pvalue = 0.4742 \nleq \alpha(0.05)$, we fail to reject H_0 and the cans are properly labeled since there is no significant difference.

Doing the test with summary statistics rather than data:

```
xbar=mean(cans); s=sd(cans); n=length(cans)
rbind(xbar,s,n)
```

```
          [,1]
xbar 37.465000
s     2.265324
n    10.000000
```

```
mu0=38; n=length(cans)
se=s/sqrt(n)
tcalc=(xbar-mu0)/se
# pvalue for 2-tailed test when Ha is not equal
# use of abs() is absolute value to account for a negative t-score
# (just leave the abs() alone regardless of tcalc)
pvalue=2*(1-pt(abs(tcalc),df=n-1)); pvalue
```

```
[1] 0.4742219
```

Hypothesis test for one proportion

Ingots are huge pieces of metal often weighing more than 10 tons (20,000 lbs.). In one plant, only about 80% of the ingots have been defective-free. In an attempt to reduce the cracking, the plant engineers and chemists have tried some new methods for casting the ingots and from a sample of 500 ingot cast in the new method, 16% of the casts were found to be defective (cracked). Is there sufficient evidence that the defective rate has decreased?

$$H_0 : p = 0.2 \ vs. \ Ha : p < 0.2$$

```
p0=.2; q0=1-p0; n=500; phat=.16
se=sqrt(p0*q0/n); se
```

```
[1] 0.01788854
```

```
zcalc=(phat-p0)/se; zcalc
```

```
[1] -2.236068
```

```
# pvalue for lower tail test (Ha: <)
pvalue=pnorm(zcalc); pvalue
```

```
[1] 0.01267366
```

We will reject H_0 if $pvalue \leq \alpha(0.05)$. Since $pvalue = 0.0217 \leq \alpha(0.05)$, we reject H_0 so the defective rate has decreased (the new method is effective).