Topic 2: Sampling distribution of the mean

Background for inference of location parameter in location/scale families

Sonja Petrović
Created for ITMD/ITMS/STAT 514

Spring 2021.

Context

Statistics:

- Functions of random variables
- Therefore, are random variables themselves.
  - In particular, they have their own distributions, called sampling distributions.
  - Meaning of sampling distribution **(Review)**
- How does inference relate to analytics? **(Review)**

Context

Statistics:

- Functions of random variables
- Therefore, are random variables themselves.
  - In particular, they have their own distributions, called sampling distributions.
  - Meaning of sampling distribution **(Review)**
- How does inference relate to analytics? **(Review)**

Do *you* know population variance?

Recall the random variable which has a normal distribution by the CLT:

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}.$$

Do *you* know population variance?

Recall the random variable which has a normal distribution by the CLT:

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}.$$

What happens when you do not know σ?

Do *you* know population variance?

Recall the random variable which has a normal distribution by the CLT:

$$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}.$$

What happens when you do not know σ?

Question

What do we know about

$$\frac{\overline{X} - \mu}{S / \sqrt{n}}?$$

A comparative simulation: Z

Suppose we just take one sample, X, of size 100 from a normal population with mean $\mu = 25$ and standard deviation $\sigma = 10$. Here's our friend, the random variable Z:

```
get.Z.value <- function(sample.size,mu,sigma){
  x <- rnorm(n=sample.size,mean=mu,sd=sigma)
  (mean(x)-mu)/(sigma/sqrt(sample.size))
}
```

Notice a **function** up there, in the code?

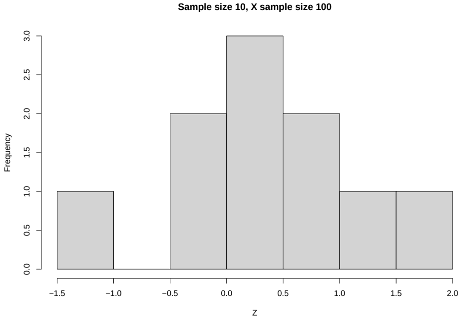
```
my.mu = 25
my.sigma = 10
my.sample.size = 100
get.Z.value(sample.size=my.sample.size,mu=my.mu,sigma = my.s
```

```
[1] -0.4039858
```

A comparative simulation: Z

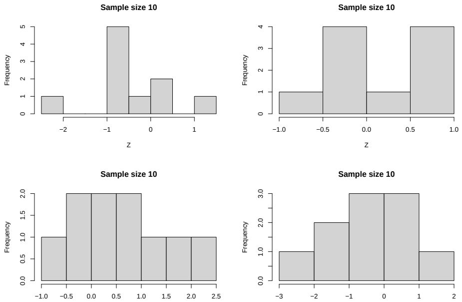
Of course this is just *one* value of the staistic Z. Now, repeat!

```
Z.sampled <- replicate(10,
                       get.Z.value(sample.size=my.sample.size,
                                   mu=my.mu,sigma = my.sigma))
hist(Z.sampled,
     main=paste("Sample size 10, X sample size",my.sample.size),
     xlab="Z")
```



A comparative simulation: Z

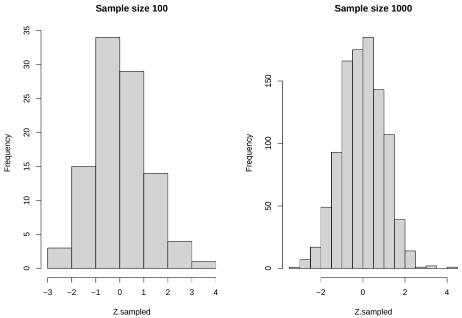
```
par(mfrow = c(2, 2))
hist(replicate(10,get.Z.value(sample.size=my.sample.size,mu=my.mu,
                  sigma =my.sigma)),main="Sample size 10",xlab="Z")
hist(replicate(10,get.Z.value(sample.size=my.sample.size,mu=my.mu,
                  sigma =my.sigma)),main="Sample size 10",xlab="Z")
hist(replicate(10,get.Z.value(sample.size=my.sample.size,mu=my.mu,
                 sigma =my.sigma)),main="Sample size 10",xlab="Z")
hist(replicate(10,get.Z.value(sample.size=my.sample.size,mu=my.mu,
                  sigma =my.sigma)),main="Sample size 10",xlab="Z")
```



A comparative simulation: Z

... What about more reps?

```
par(mfrow = c(1, 2))
Z.sampled <- replicate(100,get.Z.value(
  sample.size=my.sample.size,mu=my.mu,sigma = my.sigma))
hist(Z.sampled, main="Sample size 100")
Z.sampled <- replicate(1000, get.Z.value(
  sample.size=my.sample.size,mu=my.mu,sigma = my.sigma))
hist(Z.sampled, main="Sample size 1000")
```



A comparative simulation: T

Now let's replace σ by S.

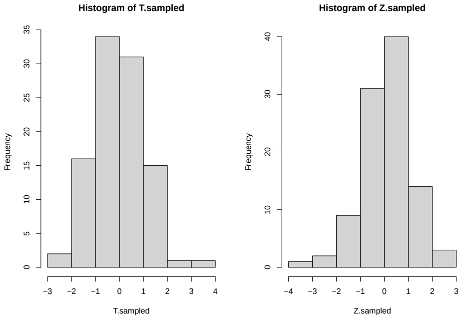
```
get.T.value <- function(sample.size, mu, sigma){
  x <- rnorm(n=sample.size,mean=mu,sd=sigma)
  (mean(x)-mu)/(sd(x)/sqrt(sample.size))
}
my.mu = 25
my.sigma = 10
my.sample.size = 100
get.T.value(sample.size=my.sample.size,mu=my.mu,sigma=my.sig
```

```
[1] -0.7673518
```

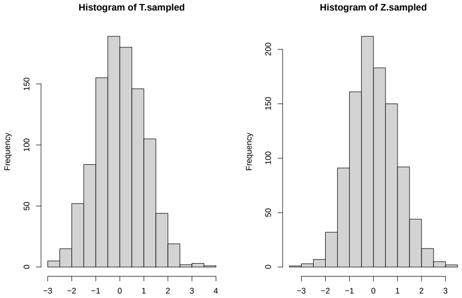
A comparative simulation: T

Again, need to repeat:

```
par(mfrow = c(1, 2))
T.sampled <- replicate(100, get.T.value(sample.size=my.sample.size,r
hist(T.sampled)
Z.sampled <- replicate(100, get.Z.value(sample.size=my.sample.size,r
hist(Z.sampled)
```



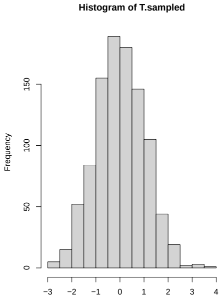
A comparative simulation: the Z vs. T random variable



Aha!

…What do you see??

A comparative simulation: the Z vs. T random variable



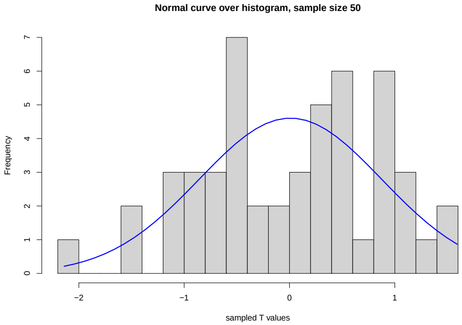
Aha!

...What do you see??

- The T distribution looks just like Z. Let's use the CLT! $\leftarrow$??
- I don't get it. We just replaced σ by S so...big deal?? $\leftarrow$??
- T is a different random variable than Z. The histograms look similar... but can we prove it's the same shape?? $\leftarrow$??

How do we compare the two sampling distributions?

One way: let us plot the T histogram and overlay the the Z density curve on top.



*For HW, you will explore different ways to plot a normal curve over a histogram.

Why?

Because we know the theoretical distribution of Z!

The t distribution

Theorem

Let $X_1, \ldots, X_n$ be independent random variables that are all normal with mean μ and standard deviation σ. Let $\overline{X}$ and $\S^2$ be sample mean and sample standard deviation, respectively. Then Then the random variable

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

has a t-distribution with $\nu = n - 1$ degrees of freedom.

- What are 'degrees of freedom' and how does $n - 1$ change the shape of $T \sim t_{n-1}$?
- How is this distribution defined?
  $\{$(By the way, have you ever wondered how is the normal $N(\mu, \sigma)$ defined? It's not "just a curve", there's a prob. formula, right?)$\}$

We need some standard notation.

- Discuss: the notation t_α.
- Discuss aslo: the notation z_α.

We need some standard notation.

- Discuss: the notation t_α.
- Discuss aslo: the notation z_α.

Aha!

Do we get the meaning of these?

An example that needs the t-distribution computation

Problem.

A chemical engineer claims that the population mean yield of a certain batch process is 500 grams per milliliter of raw material. To check this claim he samples 25 batches each month. If the computed t-value falls between $-t_{0.05}$ *and* $t_{0.05}$*, he is satisfied with this claim. What conclusion should he draw from a sample that has a mean* $\overline{X} = 518$ *grams per milliliter and a sample standard deviation* $s = 40$ *grams? Assume the distribution of yields to be approximately normal.*

An example with t in Python

```
from scipy.stats import t
mean=518
n=25
s=40
u=500
#computing the t-value from the sample
tvalue=(mean-u)/(s/5)
#computing the t0.5,... remember the degree of freedom is 24
interval=t.cdf(0.5,24)
print(-interval)
```

```
-0.6891856388430067
```

```
print(interval)
```

```
0.6891856388430067
```

```
print(tvalue)
```

```
2.25
```

What is next?

- The F-distribution (. . . almost the last one!)
- The sampling distribution of S^2.
- Putting it all together, and doing formal statistical tests:
  - Data scenarios
  - Flowchart: which distribution to use when and why?
  - Play with examples and test things out in lab work.

Aha!

Let's get started on some basic questions for "what to use when", shall we? :)

License