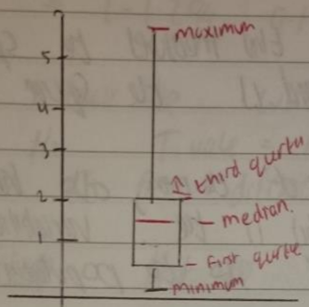
1 2013 Q2

I would use a boxplot to graph the 30 individual values for each method

Box plots are drawn in the following way



This is a simple way to represent statistical data on a plot in which a rectangle is drawn to represent the second and third quartile (50% of data), usually with a ~~vertical~~ line or horizontal line inside it representing the median

The upper and lower quartiles are the vertical lines extending above and below the box

Dots or stars above or below box represent outliers

Comparing the box plots visually will give us an indication of the possible difference between the two methods

b. ~~A~~ N = Sample Size which equals 30 for both method

Mean - For each sample the sum of the values divided by the count of values. As you can see, the means are relatively close

StDev - Standard deviation, this measures the spread of the values from the mean and is the square root of the variance

SE Mean - Standard error of the mean, also know as standard deviation of the mean, this is the variability of the mean of the sample compared to the population mean

Difference = mu (Type A) - mu (Type B)
This is the two sample testing the hypothesis that there is no difference between Type A population mean and ~~Type~~ Type B mean population mean

If population means are the same $U_1 - U_2 = 0$
H_0 : Population means the same
H_1 : Population means not the same

Estimate for difference: This is mean A - mean B which is used in the t_i-test formula.

95% CI for difference: Degrees freedom is $n_1 + n_2 - 2 = 60 - 2 = 58$
$\alpha = 0.05\%$. Using these two values we can create a t-interval where t lies above and below some value.

If the calculated t lies within this interval, we fail to reject H_0

3

equation for CI is estimator $\pm$ $t_{critical}$ se.(estimate)

$$-0.569 \pm 2.008(0.36 \qquad \frac{24}{40} \quad \frac{26}{40}$$

$$\pm\ 0.7220434967$$

$$\pm\ 0.725$$

$$= (-1.294,\ 0.156)$$

Tcalculated value must lie within interval if we are to accept H_0. T value $= -1.57$ which is outside the 95% CI for difference

We can conclude that the population means are different.

P value is also important. If p value is less than α value of 0.05 we also reject null hypothesis

The p value is the probability that the means are the same but if this probability is less than 0.05 we reject / ignore it.

c There are two types of errors type 1 and type 2

		null hypothesis true	null hypothesis false
~~accept~~ null	accept	Ok	Type 2
	reject	type 1	ok

Type 1 error is the significance level α which we set ourselves. Probability of rejecting the null hyp when it is true

Type 2: error is probability of accepting null hypothesis when it is false

Type 1 error

The probability of a type 1 error is the level of significance of the test hypothesis denoted by alpha α.

example: If chloesterol level of a healthy man is normally distributed with mean 180 and St.20, and men with levels over 225 are diagnosed as not healthy, what is probability of type 1 error:

$z = \frac{225-180}{20} = 2.25$, corresponding t-tail area is 0.0122

- The larger a sample size, the more likely a hypothesis test will detect a small difference
- When we are setting $\alpha = 0.05$ what we are actually saying is

Type 2 error.

Occurs when we reject alternate hypothesis when AH is true. This probability is denoted by "beta" β.

example: if By using larger value of n we will get lower $t_{critical}$ values which will lead to a more accurate results

The "Power" test is used to calculate probability of type II.

The sample size determines the amount of sampling error inherent in a test result. Other things equal, effects are harder to detect in smaller sample sizes. Increasing sample size is easiest way to boost statistical power of the test.

ERRORS TYPES

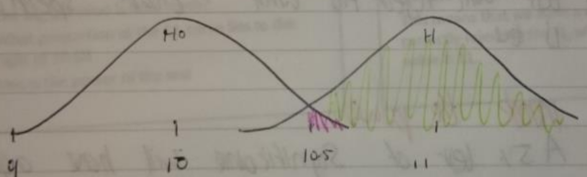
We create a hypothesis
H_0 long term mean $= 10$
H_1 long term mean $\neq 10$.

We chose a different value for mean say 11.
Assume SD = 1.5.

What would you expect to happen when you took samples of size 21?

We look at the overlap of the sampling distributions under the H_0 and H_1 (11cm)

Create the 95% CI using t dist and n-1 d.f.



The two distributions overlap. But by how much?
What proportion of the 11 curve lies to be right of 10.68 (95% for 10)
This is the power of the test

$t = \frac{10.68 - 11}{0.327} = 0.98$

= 83%

Means we have a 83% chance of correctly rejecting the H_0 when the long term value is 11

- We can increase this chance by changing the value of n

A larger n will give a smaller S.E and thus a larger t value
larger sample size yields higher power

Statistical power is the probability of correctly rejecting a false null hypothesis when a specific alternative hypothesis is true

Power test influenced by
- Difference between actual population mean and null hypothesised mean
- variability of data σ
- sample size N
- alpha error α

The power against a specific alternate is calculated as the probability that the test will reject Ho when that specific alternate is true

Ways to increase the power
- Increase α. A 5% test of significance will have a greater chance of rejecting the alternate than a 1% test because the strength of the evidence required for rejection is less
- Consider a particular alternative that is farther away from the value of μ than are in Ho but the closer to the hypothesised value μ_0 are harder to detect than values of μ that are farther from μ_0
- Increase sample size. More data will provide more info about $\bar{x}$ so we have a better chance of distinguishing values of μ
- Decrease σ. This has same effect as increasing sample size, it provides more info about μ. Improving the measurement process and restricting attention to a subpopulation are two common ways to decrease σ.

Stats First Year

Change to standardised normal mean 0 st. d 1 by

$$Z = \frac{X - \mu}{\sigma}$$

For Samples $\sigma = \frac{\sigma}{\sqrt{n}}$ also called Standard Error.
By central limit theorm

Sampling proportions

$p = \frac{20}{80} = 0.25$

$SE(p) = \sqrt{\frac{p*(1-p)}{n}} = 0.048$

$95\%\ CI = \hat{p} \pm 1.96\, SE(p)$
$0.25 \pm 1.96(0.048)$

$n^* p^*(1-p)$ has to be >5 to work

Deciding how wide our CI will be

$\hat{p} \pm 1.96 \times \sqrt{\frac{p(1-p)}{n}}$

we decide on a given error say 0.02

$$1.96 \times \sqrt{\frac{p(1-p)}{n}} < 0.02$$

manipulate get us: $n > \frac{\hat{p} \times (1-p) \times 1.96^2}{0.02^2}$

In worst case $p = 0.50 \Rightarrow$ sub in to get value

For Sample Sizes for estimating means

$95\%\ CI = \bar{x} \pm 1.96 \times \frac{Sd}{\sqrt{n}}$

d is the width

$$n \geq \frac{1.96^2 \times Sd^2}{d^2}$$

Test statistic for sample and hypothesis

$$z = \frac{\mu_0 - \text{sample mean}}{\frac{sd}{\sqrt{n}}}$$ μ_0 is population mean

Testing population mean v sample mean

Example - population mean = 250
sample = 248.42
sample sd sd
8.98 = sd sample

$$\frac{250 - 248.42}{\frac{8.98}{\sqrt{50}}} = 1.26$$

Or we could use our z value and see if it lies within (-1.96, 1.96) interval, accept if within

Or calculate CI by sample mean ± 1.96 * SE, if population mean within CI accept H_0

2

T-test - two sided

two group $t = \frac{(\bar{x}_{new} - \bar{x}_{normal}) - \text{hyp val}}{se(\bar{x}_{new} - \bar{x}_{normal})}$

$$SE[\bar{x}_1 - \bar{x}_2] = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

or poold se. $= s \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

$$s^2 = \frac{(n_1 - 1) \times s_1^2 + (n_2 - 1) \times s_2^2}{n_1 + n_2 - 2}$$

95% for CI: (2, 3)
bounce for new process is betw 2 and 3cm higher than normal proces

ONE SIDED

Is the population (long term mean bounce height) from new process higher/lower than the usual proces

H_0: $\mu_{normal} \le \mu_{new}$ or $\mu_{new} - \mu_{normal} \le 0$

H_1 $\mu_{normal} > \mu_{new}$ or $\mu_{new} - \mu_{normal} > 0$

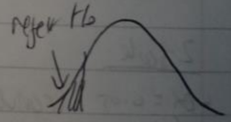
value $\ge$ critical t, evidence against H_0

H_0: $\mu \ge 750$ hr

H_1: $\mu < 750$ hr

critical $t = -1.69$

value < -1.69 evidence against H_0



If we want 5% we use two tailed but $\alpha = 10\%$ to get critical t

1 STATS QUICK NOTES

$p \pm 1.96 * \sqrt{\frac{p(1-p)}{n}}$ for sample proportion

$1.96 * \sqrt{\frac{p(1-p)}{n}} < 0.02$

$1.96^2 \left(\frac{p(1-p)}{n}\right) < 0.02^2$

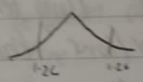
$\frac{1.96^2 \, p(1-p)}{0.02^2} = n$ $n = 2401$

$p \pm 1.96 \frac{st}{\sqrt{n}}$ for means

$z = \frac{\mu_0 - \text{sample mean}}{\frac{st}{\sqrt{n}}}$ t-test. $\frac{\mu_0 - \text{sample mean}}{\text{s.e.}}$

Testing - 3 ways

1. Calculate Stand Error



2. Calculate probability of tail.
   $2[1 - P(z < 1.26)]$ $= 0.2$
3. Probability of 0.2 of getting our data
4. Compare to $\alpha = 0.05$, if bigger accept H_0.

Z-Scale

- $\alpha = 0.05$ convert to $z \Rightarrow$ ± 1.96
- Calculate z, if in $(-1.96, 1.96)$ accept H_0
  1.27 is ✓

CI

value ± 1.96 s.e. $248.40 \pm 1.96 * 1.27$ $(245.91, 250.89)$

if H_0 value in interval, accept ✓

3

CHI-SQUARE

Qmyra.	June	December	
Yes	50	60	110
No.	450	440	890
Total	500	500	1000

H_0: no association between the two variables (whether you listen to Qmyra) and time (June vs. December) in the population

i.e. Proportion of people in population who listen in June in population is the same as proportion who listen in december in population.

$\pi_{June} = \pi_{December}$

H_1: There is an association between the two variables, whether you listen to Qmyra and time.

i.e. Proportion of people who listen in June in population is not same as proportion who listen in December in population

$\pi_{June} \neq \pi_{Decem}$

Expected value

$\frac{\text{row total} * \text{column total}}{N}$

Qmyra.	June	December	observe	
Yes	55	55	50	60
No.	445	445	450	440

How To Compare

$$\Sigma \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

- Will be big if observed values differ from expected

4

χ^2 calc

$$\sum \frac{(O-E)^2}{E} = \frac{(50-55)^2}{55} + \frac{(60-55)^2}{55} + \frac{(450-445)^2}{445} + \frac{(440-445)^2}{445}$$

$= 0.455 + 0.455 + 0.056 + 0.056 = 1.021$

df = (number rows − 1) * (number columns − 1)
(2−1) * (2−1) = 1

From table $\alpha = 0.05$, df = 1
critical value = 3.84

- Value > 3.84 − evidence against H_0
- Value ≤ 3.84 − not enough evidence against H_0
- 1.021, no evidence against H_0

CI for difference in Proportion

$(P_{Dec} - P_{June}) \pm t_{critical} * SE(P_{Dec} - P_{June})$

$n_1 + n_2 - 2$ df = 500 + 500 − 2 = 998 ⇒ 1.96

$$SE(P_1 - P_2) = \sqrt{\frac{P_1 * (1-P_1)}{n_1} + \frac{P_2 * (1-P_2)}{n_2}}$$

$$\sqrt{\frac{0.12 * (1-0.12)}{500} + \frac{(0.10)(1-0.1)}{500}} = 0.02$$

$0.02 \pm 1.96 * 0.02 = 0.02 \pm 0.04 = (-0.02, 0.06)$

0 in interval, no evidence against difference, accept H_0

5.

Chi-Square

- Chi-Square does not give any info about strength of relation
- Only conveys the existence / non existence of relationship between the variables investigated

Power

	Null hypothesis True	Null hypothesis False
Accept	OK $p = 1-\alpha$	Type II error $p = \beta$
Reject	Type I $p = \alpha$	Correct $p = 1-\beta$

Type II probability of accepting a false H_0

Power is probability of rejecting when H_0 is false
(1 - Type 2 error)

repeat for 11,12,13 for jack queen king

```
if (decknamed == "11"){
    decknamed [2,i] = "A"
}

~~example~~ shuffled.deck.idx = sample(1:52, size=52, replace=FALSE)

shuffled.deck = decknamed [ , shuffled.deck.idx]

# store card.
N=1000
first.card = matrix(nrow=N, ncol=2)

for (j in ~~1:1000~~ 1:N){
    shuffled.deck.idx = sample(1:52, Size=52, Replace=FALSE)
    ~~first.card = [j, ] = deck.named~~ [ , shuffled.deck [.
shuffled deck = deck named [ , shuffled.deck.idx]
    first card [j, ] = shuffled.deck
}
             length(                [ ,1]
no.queens = which(first.card == "Queen"))
```

5/02/15 STATS

"\t" tab
"\n" new line

```
For(j in 1:100)
{
  cat("\n", j)
}
```
prints 1 to 100

```
for(j in 1:20) {
Str = paste("Number", j)
print(Str)
}
```
concantenate "number" and j

x = rnorm(x, mean, sd) rnorm(5, 10, 1) e.g

qnorm - quantile - point on axis with x% probability below it.
pnorm - percentile - cumulative
dnorm - density

x = runif(100) - uniform distribution 0,1

Gamma distribution α β (shape, rate) $f(x) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$

Set $\alpha = 1$ we get exp distribution $f(x) = \beta e^{-\beta x}$

```
#shuffling a deck of cards
deck = 1:52
Shuffled = Sample(deck, size=52, replace=FALSE)
```
Sample - shuffle(s) anything / permutate it.

```
Suit = c(rep('spades',13), rep('clubs',13), rep('hearts',13), rep('diamonds',13))
#gives each suit named out 13 times
```

```
denominator = c(1:13, 1:13, 1:13, 1:13)
decknamed = rbind(Suit, denomination)
```
cbind for columns

```
For (decknamed i in 1:52)
  if (decknamed[2, i] == "1"){
    decknamed[2, i] = "ace"
  }
```
Second row → the number