Assumptions

Can we draw accurate conclusions from our tests?

Assumptions

Most statistical tests make assumptions about your data

When these assumptions are broken, the tests are no longer accurate!

My goal:

Teach you about the assumptions of a test

My approach:

- Show you how to test whether your data meets the most common assumptions
- Suggest fixes for when these assumptions are not met

Assumptions

Assumptions of parametric tests:

- Normally distributed data
- Homogeneity of variance
- Interval data
- Independence

Normality

Does my data follow the normal distribution?

Normality

Technically, the error distribution should be normal

In most linear models, this means that the sampling distribution of our outcome value should be normal

Why? Because outcome = (model) + error; model is fixed, so if the sample is normal, then the error is normal

We don't know the sampling distribution, so we look at the sample itself

If a value is normally distributed within the sample, then the statistic (e.g. mean) is normal between samples as well

Histogram check

Read the data

File: DownloadFestival.dat, set Name to festivalData

Dataset: festival-goer hygiene (repeated measures)

Variables:

ticknumb: participant id

gender: male/female

day1, day2, day3: hygiene level at days 1-3 (0-4 scale)

Histogram check

Fix the outlier

```
festivalData[festivalData$day1 == 20.02]$day1 <- 2.02
```

Create a density histogram for the data at day1:

```
histo <- ggplot(festivalData, aes(day1)) +
geom_histogram(aes(y=..density..), color="black",
fill="white") + labs(x = "Hygiene at day 1", y = "Density")
```

Histogram check

Plot a normal curve over the plot with the same mean and standard deviation as our data:

```
histo + stat_function(fun = dnorm, args = list(mean =
mean(festivalData$day1, na.rm=T), sd =
sd(festivalData$day1, na.rm=T)))
```

dnorm plots a normal distribution, takes arguments “mean” and “sd”

Need to remove missing values (na.rm=T)

Try it yourself: do the same for days 2 and 3 (no outliers)

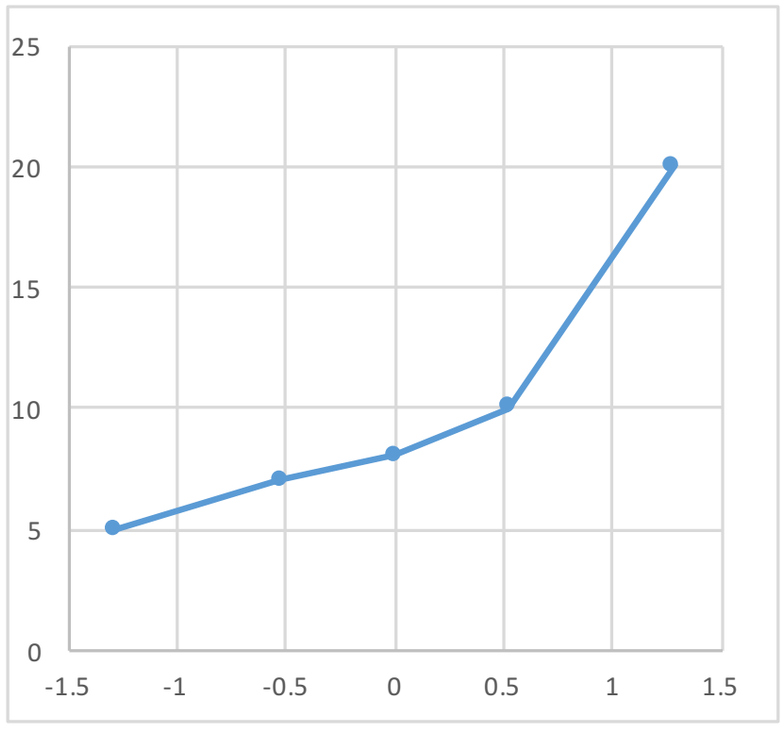
QQ-plot check

Sort the data, then plot it against the values of an ideal normal distribution

Data: 5, 7, 8, 10, 20

Percentiles: 10%, 30%, 50%, 70%, 90%

Z-scores: -1.28, -0.52, 0, 0.52, 1.28



Plot should look like a diagonal line

In R, use the qplot function:

```
qplot(sample = festivalData$day1, stat="qq")
```

Intermezzo

Useful Z-scores:

$p < .05$ for $-1.96 < z < 1.96$

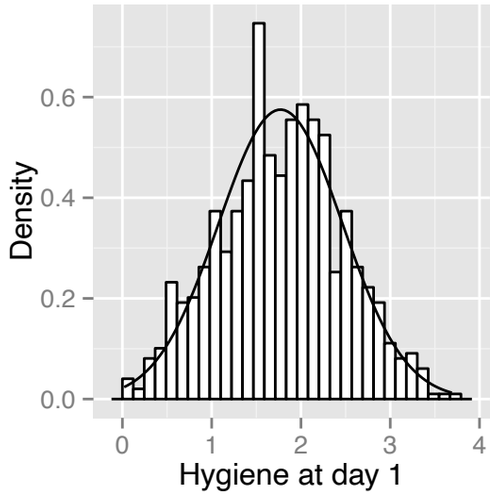
$p < .01$ for $-2.58 < z < 2.58$

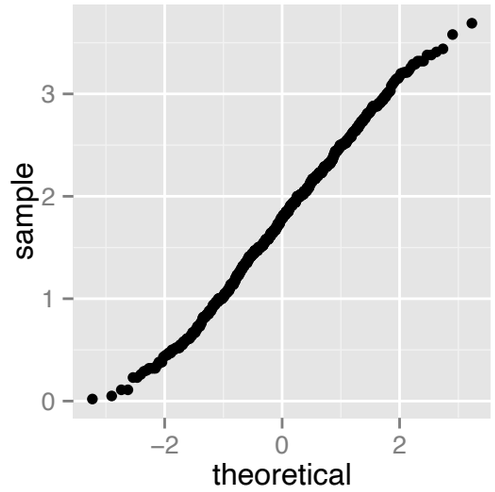
$p < .001$ for $-3.29 < z < 3.29$

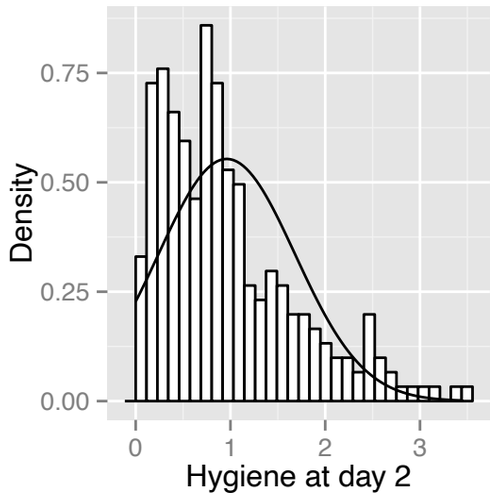
Evaluate

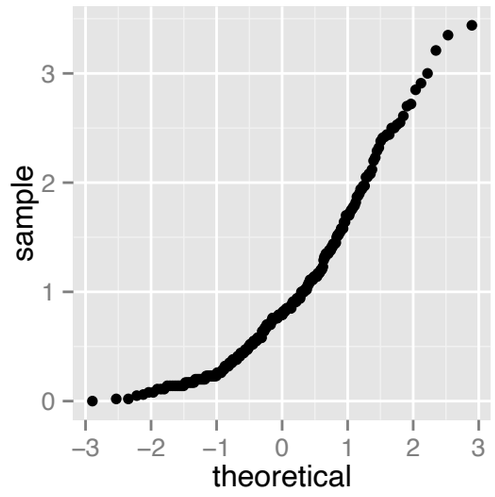
Which data look normal?

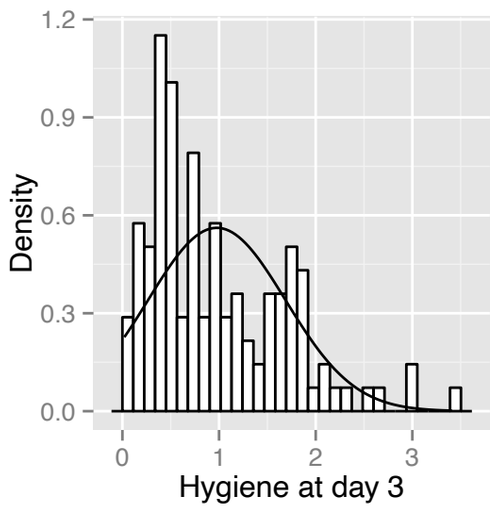
Which don't?

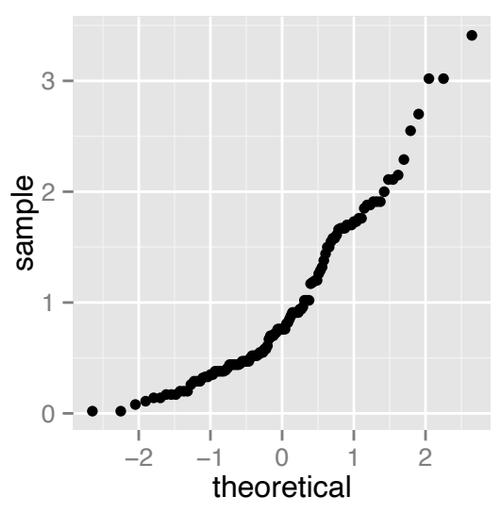












Numbers check

Install *pastecs* package

Report normality statistics, using stat.desc:

round(stat.desc(festivalData[,c("day1","day2","day3")], basic = F, norm = T), digits = 3)

We ask for all three days at the same time (select columns day1, day2, and day3), and round to 3 digits (easier to read)

Look at Skewness and Kurtosis and their z-scores

Z_S = skew / SE_{skew} and Z_K = kurt / SE_{kurt}

Numbers check

	day1	day2	day3
median	1.790	0.790	0.760
mean	**1.771**	**0.961**	**0.977**
SE.mean	0.024	0.044	0.064
CI.mean.0.95	0.048	0.087	0.127
var	0.481	0.520	0.504
std.dev	0.694	0.721	0.710
coef.var	0.392	0.750	0.727
skewness	**−0.004**	**1.083**	**1.008**
skew.2SE	−0.026	3.612	2.309
kurtosis	**−0.422**	**0.755**	**0.595**
kurt.2SE	−1.228	1.265	0.686
normtest.W	0.996	0.908	0.908
normtest.p	0.032	0.000	0.000

Z_S = -0.052 (d1), 7.224 (d2), and 4.609 (d3)

Z_K = -2.456 (d1), 2.330 (d2), and 1.372 (d3)

Different groups

Read the data

File: RExam.dat, set Name to rexam

Dataset: R exam scores at two universities

Variables:

exam: exam score

computer: computer literacy

lectures: percentage of lectures attended

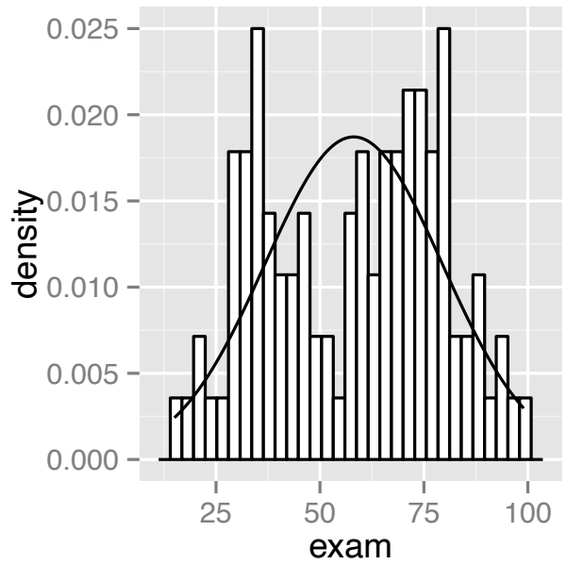
numeracy: numeracy score

uni: university (0 = Duncetown, 1 = Sussex)

Different groups

Draw histogram for exam score with normal curve:

```
ggplot(rexam,aes(exam)) + geom_histogram(aes(y = ..density..),
color="black", fill="white") + stat_function(fun = dnorm,
args=list(mean = mean(rexam$exam), sd = sd(rexam$exam)))
```



Looks bimodal… What about per university?

Different groups

Numbers per university: use “by” function

```
by(rexam$exam, rexam$uni, stat.desc, basic=F, norm=T)
```

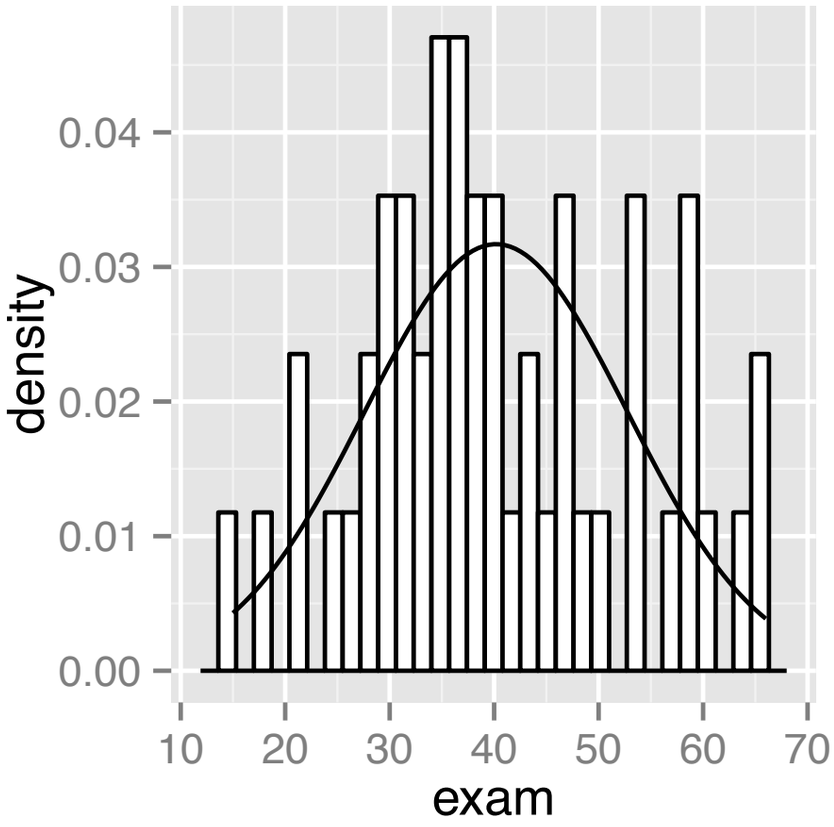
Plot per university: subsetting

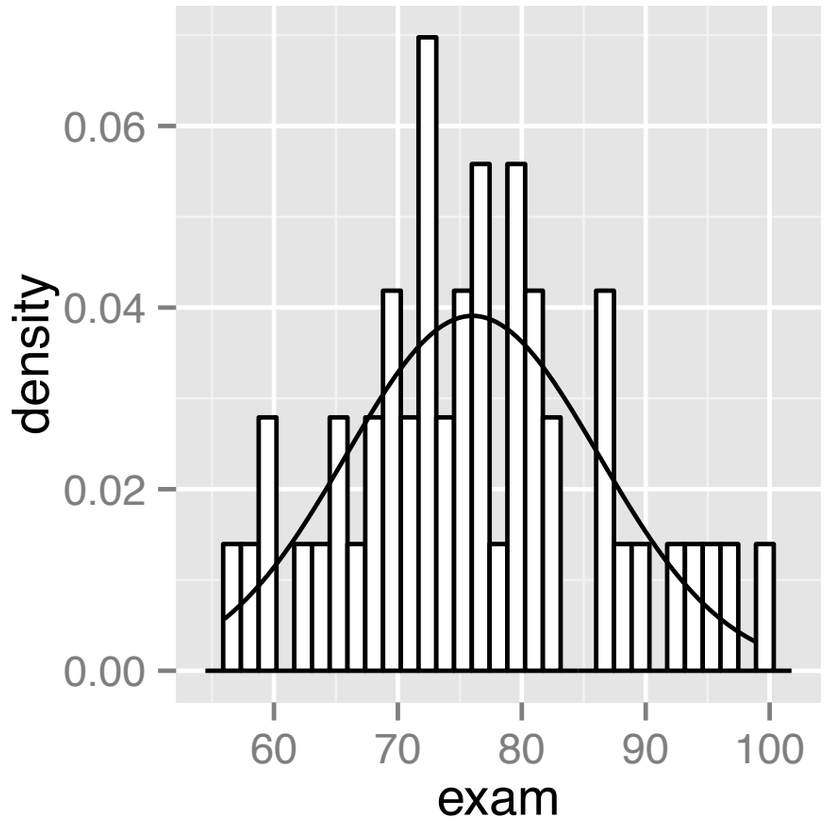
```
dunceData=subset(rexam,rexam$uni==0)

sussexData=subset(rexam,rexam$uni==1)

ggplot(dunceData,aes(exam))+
geom_histogram(aes(y=..density..), color=“black”,
fill=“white”) + stat_function(fun = dnorm, args = list(mean =
mean(dunceData$exam), sd = sd(dunceData$exam)))
```

Different groups





Lesson: normality may depend on the model!

If we test the effect of university (our model), the errors are actually normal, because we model the bimodal means

Testing normality

Shapiro–Wilk test

Has a test statistic *W*, and a *p*-value

$p < .05$ —> distribution is significantly different from 1

Warning: too sensitive for large *N*

Also, remember our lesson!

In R:

last two lines of stat.desc(var, basic=F, norm=T)

shapiro.test(var)

Testing normality

Example: RExam:

```
shapiro.test(rexam$exam)
```

```
W = 0.9613, p = .004991
```

How to report this result:

> "The R exam score, *W* = 0.96, *p* = .005, is significantly non-normal."

Testing normality

By university:

```
by(rexam$exam, rexam$uni, shapiro.test)
```

Dunce: W = 0.9722, p = .2829

Sussex: W = 0.9837, p = .7151

Remember our lesson!

Normality

Which method should I use?

For a large N:

Tests are too sensitive (a significant deviation does not have to be a substantial deviation!)

Look at the plots (overall, per group)

For a small N:

Plots are too sparse (even a substantial deviation is not necessarily significant!)

Look at the tests (Z_S, Z_K and Shapiro-Wilk)

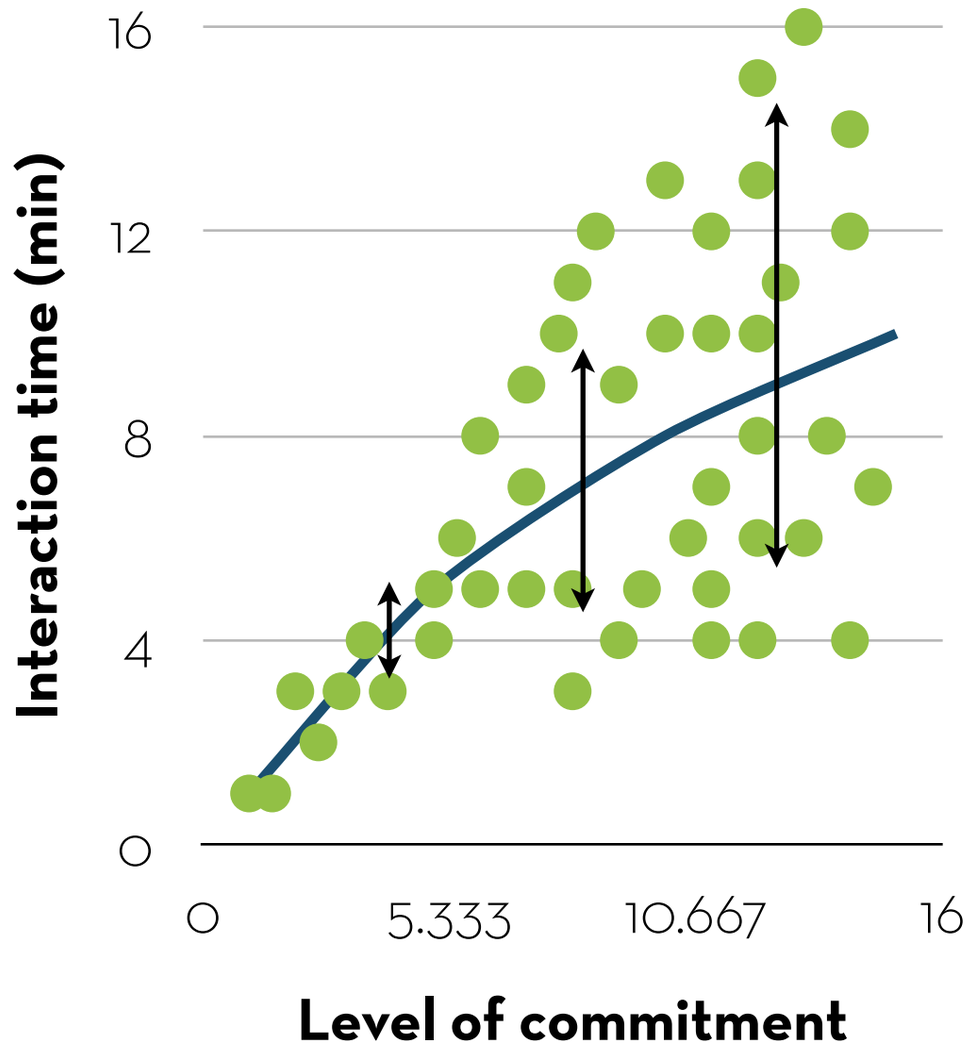
Homogeneity

For differences between groups (e.g. system A vs system B):

Is the variance (SD) the same for each group?

For continuous X -> Y:

Is the variance of Y stable at all levels of X?



Levene's test

How to test whether variance is homogeneous?

For continuous X -> Y: graphs (more on this later)

For differences between groups: Levene's test

If $p < .05$ —> heterogeneity

Warning: too sensitive for large *N*

Example

Load “car” package

Run Levene’s test on exam score per uni:

leveneTest(rexam$exam, rexam$uni)

$F(1,98) = 2.0886$, $p = .1516$

Run Levene’s test on exam score per uni:

leveneTest(rexam$numeracy, rexam$uni)

$F(1,98) = 5.366$, $p = .02262$

Example

Report:

"For the R exam score, the variances were similar for Duncetown and Sussex students, $F(1,98) = 2.09$, $p = .15$"

"For the numeracy scores, the variances were significantly different in the two groups, $F(1,98) = 5.37$, $p = .023$"

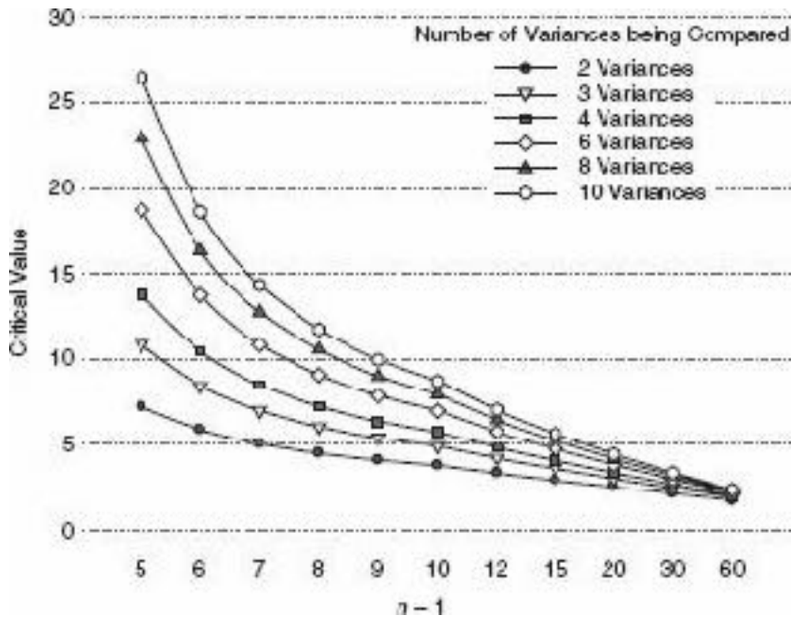
Variance ratio

Hartley's F_{max} (variance ratio): largest/smallest variance

var(sussexData$numeracy)/var(dunceData$numeracy)

Is it lower than the critical value (depends on number of groups, and *n* per group)?

If yes, then homogeneous; if no, then heterogeneous



Fixing problems

A z-score of ±3.29 has a probability of .001

What to do with such outliers?

- Remove the score (only if you have good reasons to)
- Transform the data (see next slides)
- Replace the score (with the next highest score + 1, or mean + 3.29*sd)
- Use a robust test (will be discussed later)

Remove outliers

Use the “ifelse” function

```
festivalData$day1NoOutlier <-
    ifelse(festivalData$day1 > 4, NA, festivalData$day1)
```

Transformations

log transform:

```
festivalData$logday1 <- log(festivalData$day1 + 1)
(we use +1, because log(0) does not exist!)
```

square root transform:

```
festivalData$sqrtday1 <- sqrt(festivalData$day1)
```

reciprocal transform:

```
festivalData$recday1 <-1/(festivalData$day1 + 1)
```

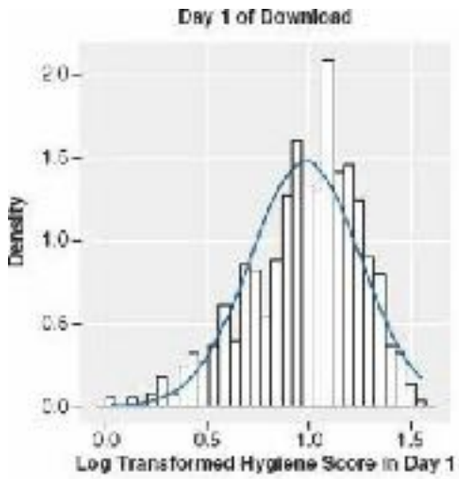
Transformations

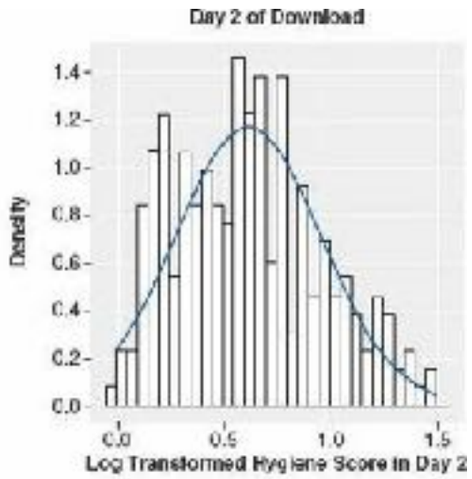
Do the same for day2 and day3

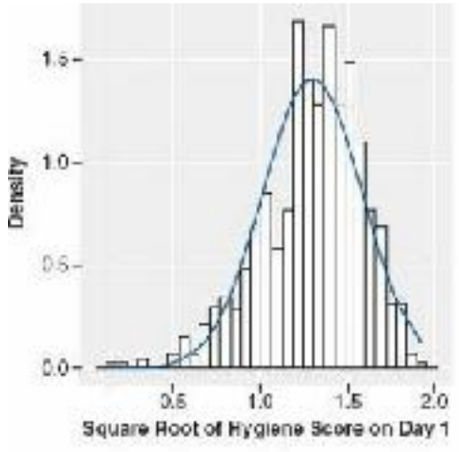
Draw histograms with normal line, e.g.:

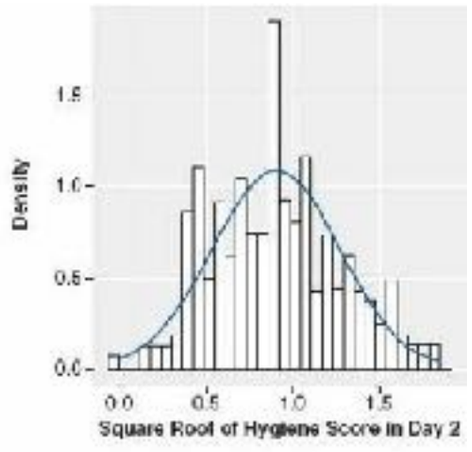
```
ggplot(festivalData,aes(logday1)) +
geom_histogram(aes(y=..density..), color = “black”,
fill=”white”) + stat_function(fun = dnorm, args = list(mean =
mean(festivalData$logday1), sd =
sd(festivalData$logday1)))
```

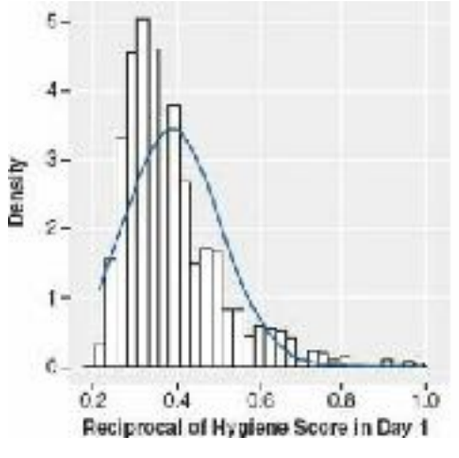
Transformations

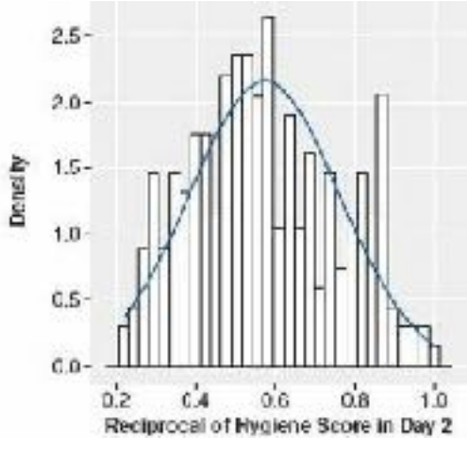












Power Analysis

for user experiments

Power Analysis

My goal:

Teach how to scientifically decide whether a sample size is sufficient for a certain study

My approach:

- Quick review of effect sizes, p-values, and power
- Intro to power analysis
- Demo using G*Power

A quick review

of effect sizes, p-values, and power

A quick review

Is my new system (version B) better than version A?

Experimental hypothesis: H1: Mb > Ma

Calculate the means. Do they differ a lot?

Given no effect, we expect the means to be roughly equal

H0: Mb = Ma

To test H1, we try to reject H0

...if the difference between Mb and Ma is so large that H0 is unlikely

P-value

P-value: likelihood that an effect this size is due to chance

- probability of this difference or larger, given H0

Weighed by the standard error (SE)

- Why? Because if the SE is large, we expect larger differences under H0, but if the SE is small, we expect smaller differences under H0
- If the difference is larger than expected based on the SE, we reject H0 (and thus, H1 is supported)

P-value depends on sample size (Why? SE depends on N!)

Effect size

Effect size: the strength of a result

- difference between Mb and Ma
- can be standardized (dividing by sd)
- does not depend on sample size (dividing by sd, not se!)

Example

Do married men weigh more than single men?

Find 4 married men: $N_m = 4$, $Mean_m = 182$, $SD_m = 15$

Find 4 single men: $N_s = 4$, $Mean_s = 170$, $SD_s = 15$

Effect size: 12 lbs

Is this a large effect? —> Need to standardize it!

Cohen's d = $(Mean_m - Mean_s)$/pooled SD

(182–170)/15 = 0.8... this is indeed a large effect

Is it significant? No! $p = .301$

Example 2

Do married men weigh more than single men?

Find 4000 married men: $N_m = 4000$, $M_m = 177.5$, $SD_m = 15$

Find 4000 single men: $N_s = 4000$, $M_s = 176.5$, $SD_s = 15$

Effect size: 1 lb

Is this a large effect?

(177.5–176.5)/15 = 0.067... this is a very small effect

Is it significant? Yes! $p = .0014$

Reflection

Small studies (N << 100) may find medium or large effects that are not significant

Waste of resources! (unless they are pilot studies)

Large studies (N >> 100) may find very small effects that are significant

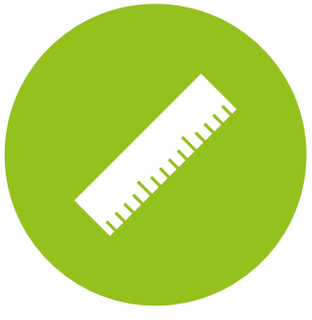
Also a waste of resources! (could have done with fewer)

How can we prevent wasting resources?

Do a power analysis!

Power analysis

an introduction



Power analysis

We reject H0 when $p < .05$

> May still be due to chance! (e.g. sample 10 men's heights repeatedly... mean will differ due to random variation)
>
> 5% of the time, two samples will be different with $p < .05$, even if they are sampled from the same population!

So, what about the 5% of the times that we reject the null hypothesis, but we got it wrong?

And what about the cases where there is a real effect but we didn't find it?

Getting it wrong

So, what about the 5% of the times that we reject the null hypothesis, but we got it wrong?

> This is a Type I error; 5% is the alpha-level

And what about the cases where there is a real effect but we didn't find it?

> This is a Type II error; we want this error to be smaller than 20%... the beta-level

Alpha and power

	There is a real effect	There is no real effect
Found an effect	**Power**	alpha (false positive)
Found no effect	beta (false negative)	1–alpha (true negative)

Power

1-beta = power

The probability of finding an effect that is really there

How high is our power? Power depends on...

...alpha (if we use $p < .01$, our power is lower)

...effect size (if the effect is smaller, power is lower)

...N (if we use a larger sample, we increase our power)

Given alpha = 0.05, and a certain expected effect size, how large should our N be to find a true effect 80% of the time?

Power analysis

A calculation involving the following 4 parameters:

- Alpha (cut-off p-value, often .05)
- Power (probability of finding a true effect, often .80 or .85)
- N (sample size, usually the thing we are trying to calculate)
- Effect size (usually the “expected effect”)

Types

A priori: compute N, given other variables

Conducted before you run your study

Post-hoc: compute power, given other variables

Conducted afterwards to find out if you had enough participants to detect a certain effect

Sensitivity: compute effect size, given other variables

Find out the minimum effect size you can detect, given the number of participants

Expected effect

An “educated guess” based on:

- Pilot study results
- Findings from similar studies
- Whatever is considered “meaningful”
- Educated guess

Examples per test

Statistic	Small	Medium	Large
Means - Cohen's d	0.2	0.5	0.8
ANOVA - Cohen's f	0.1	0.25	0.4
ANOVA - eta squared	0.01	0.06	0.14
Regression - f-squared	0.02	0.15	0.35
Correlation - r or point biserial	0.1	0.3	0.5
Correlation - R-squared	0.01	0.06	0.14
Association - 2x2 odds ratio	1.5	3.5	9
Association - w or Phi	0.1	0.3	0.5

Calculations

What if the effect size is not provided in similar studies?

Compute it!

Comparing means (e.g. t-test): Cohen's d:

$(Mean_a - Mean_b)$/pooled SD

or: $2t/\sqrt{(df)}$

Calculations

Anova: eta-squared and Cohen's f:

eta-squared = $(f)^2$ = SSm/SSt

Note: SPSS reports the *partial* eta-squared!

can also be used, but different (difficult) calculation

Regression: f-squared:

f^2 = partial R^2/(1-partial R^2)

or: calculate from ANOVA table (SSm/SSt)

We will get back to this later

G*Power demo

power analysis made easy!

G*Power demo

An existing study found that a new TurboTax interface reduced tax filing time from 3.0 hours (SD: 0.5 hours) to 2.7 hours (SD: 0.5 hours).

You created a new interface that you think is even better. How many participants do you need to find an effect that is at least the same size? (assume 85% power)

G*Power demo

You conducted a linear regression testing the effect of number of previous privacy violations on 35 Facebook users' privacy concerns (controlling for age and gender).

The number of previous violations was not significant.

The model without this variable had an R^2 of 0.15.

The model with this variable had an R^2 of 0.30.

What was your power? What sample size should you use to find an effect of this size with 85% power?

G*Power demo

You want to test the combined effect of 6 text sizes and 6 background colors on text readability. You only have money for 150 study participants.

What is the maximum effect size you can find (with 85% power) for a main effects of text size and background color?

What about the interaction effect?

Would it help if you only test 2 sizes and colors?

Final thoughts...

a few warnings, and a final cool trick...

Final thoughts...

Your Mileage May Vary!

Because power cannot be 100%, there is no guarantee you will find an effect!

The effect in your study might be smaller than in previous work!

Your may need to exclude faulty/outlying participants!

Better to estimate conservatively!

Or check out the graphs to see what would happen...

Tiny samples

Be aware of tiny samples (even when they report significant results)

- Randomization doesn't work well in tiny samples
- Tiny samples fall prey to the "publication bias"
- Due to the "winner's curse", tiny samples overestimate the real effect size

These problems are worst for counter-intuitive results

- Ask your friendly neighborhood Bayesian statistician

A final cool trick...

Let's say you need to collect 150 participants...

Ugh... 3 weeks of my time!

...why not run a quick analysis after the first 50 to see if the results are significant?

That's called "p-hacking", and is not allowed

Why? Because you inflate alpha by "peeking"

But what if you compensate by reducing your alpha?

That's allowed! It's called sequential analysis

A final cool trick...

After 50 participants, you do an analysis

3 options:

- No significance, low effect size (reaction: abandon study)
- Significant result (reaction: stop study, take 2 weeks off)
- No significance, but decent effect size (reaction: continue collecting data)

See http://dx.doi.org/10.1002/ejsp.2023 for more details...

“It is the mark of a truly intelligent person
to be moved by statistics.”

T H A N K S !

George Bernard Shaw