Statistical assumptions

Methods 1, E2021 - Lecture 4
Tuesday 21/9/2021
Fabio Trecca

Attendance registration!

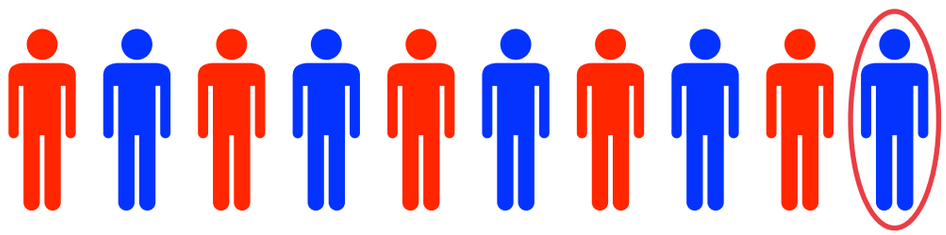

QUIZ
TIME

Quiz time (1)

- $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
- $x_i - \bar{x}$
- $SS = \sum (x_i - \bar{x})^2 = \sum (x_i - \bar{x})(x_i - \bar{x})$
- $s^2 = \frac{SS}{n-1} = \frac{\sum (x_i - \bar{x})^2}{n-1}$
- $s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$

Degrees of freedom

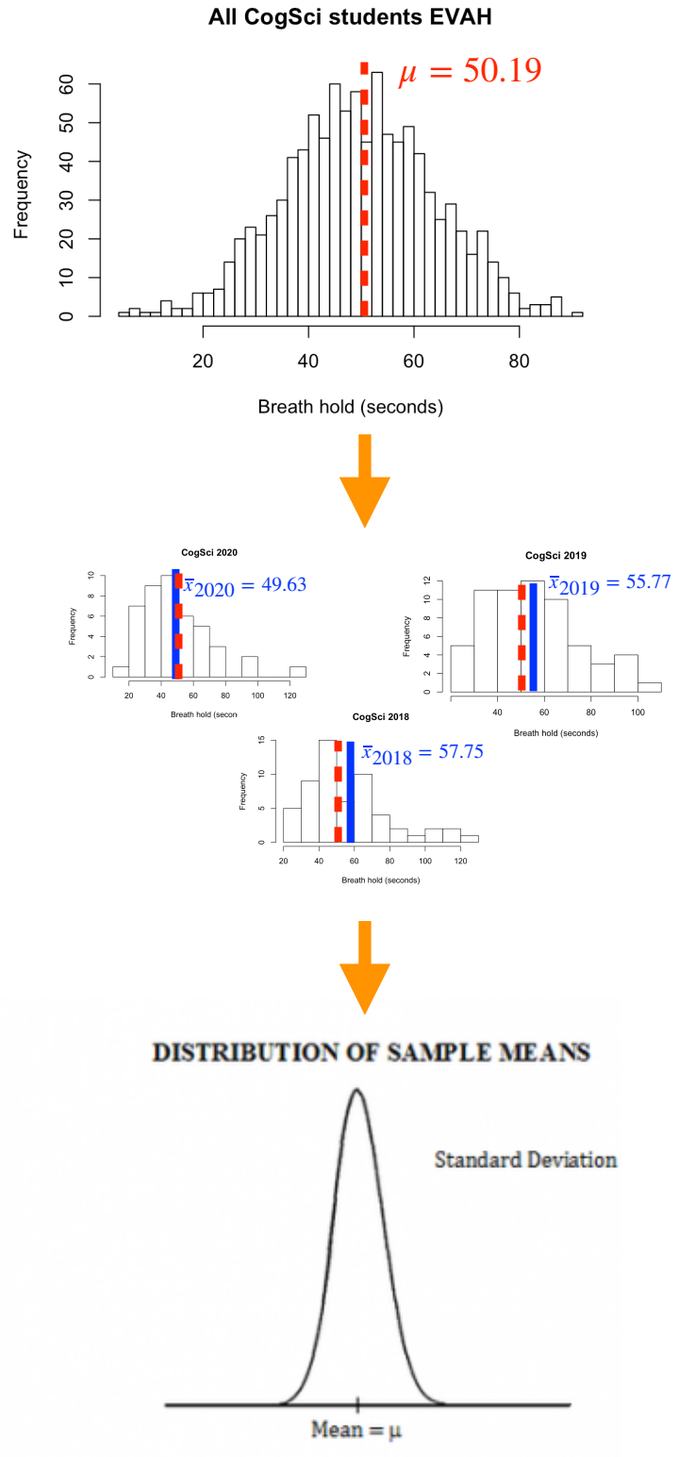
- The maximum number of logically independent values that have the freedom to vary in a data sample
- All values -1 are free to vary
- The “school yard soccer” explanation



- The “budget prioritization” explanation
  - Defense: 10%
  - Health: 65%
  - Education: 20%
  - Art: ?

Quiz time (2)

- μ, σ
- $\bar{x}, s$
- $\mu_{\bar{x}}, \sigma_{\bar{x}}$
- $\sigma_{\bar{x}}^2 = \frac{\sigma^2}{N}$
- $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}}$
- $\sigma_{\bar{x}} \approx \frac{s}{\sqrt{n}}$

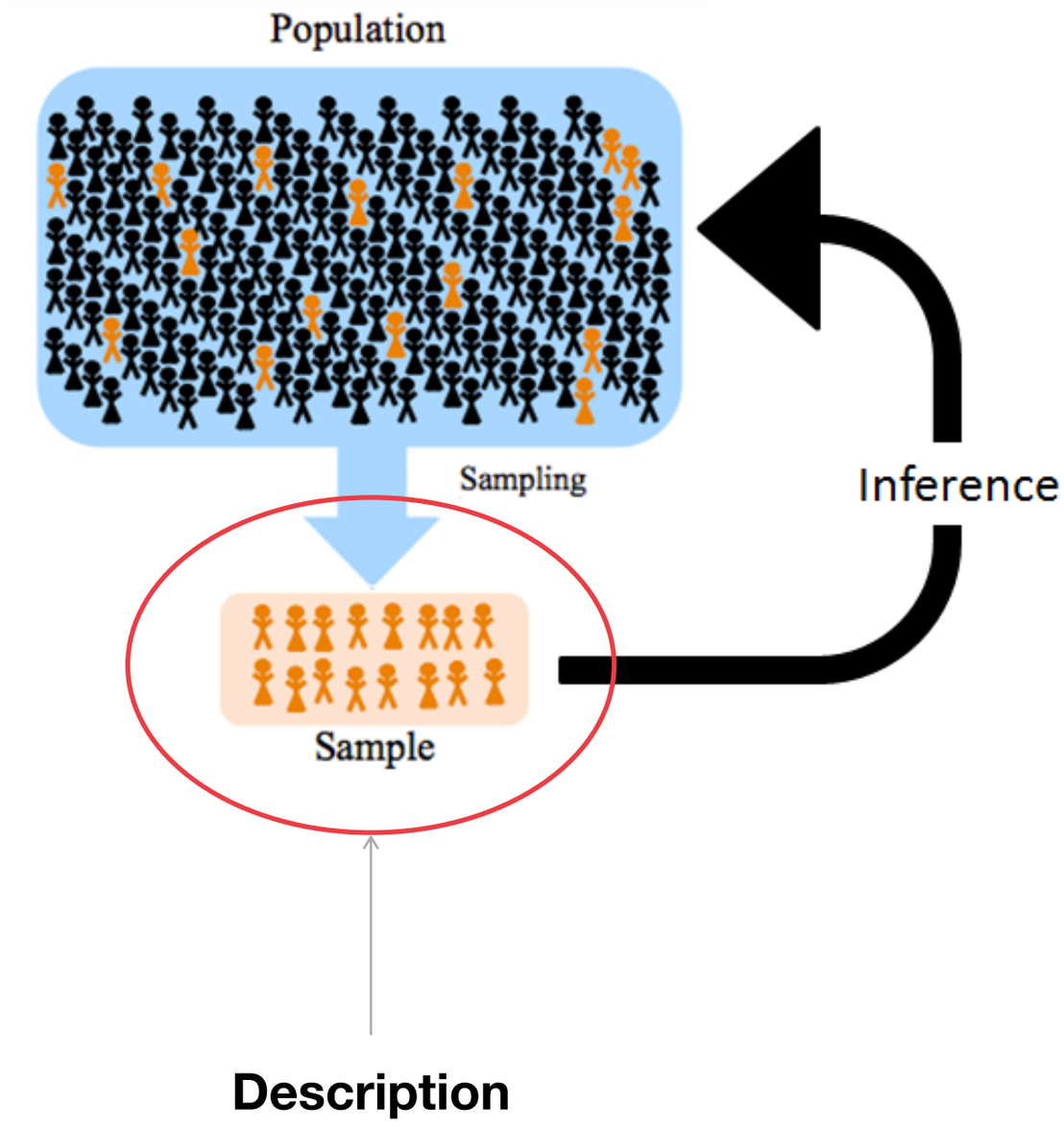


Recap: Standard Error of the Mean (1)

- A measure of the statistical accuracy of **an estimate**, equal to the standard deviation of the **theoretical distribution of a large population of such estimates**

- The sampling distribution of a mean is generated by repeated sampling from the same population and recording of the sample means obtained

- The variance of the sampling distribution obtained is equal to the variance of the population divided by the sample size (because as the sample size increases, sample means cluster more closely around the population mean — "law of large numbers")



Recap: Standard Error of the Mean (2)

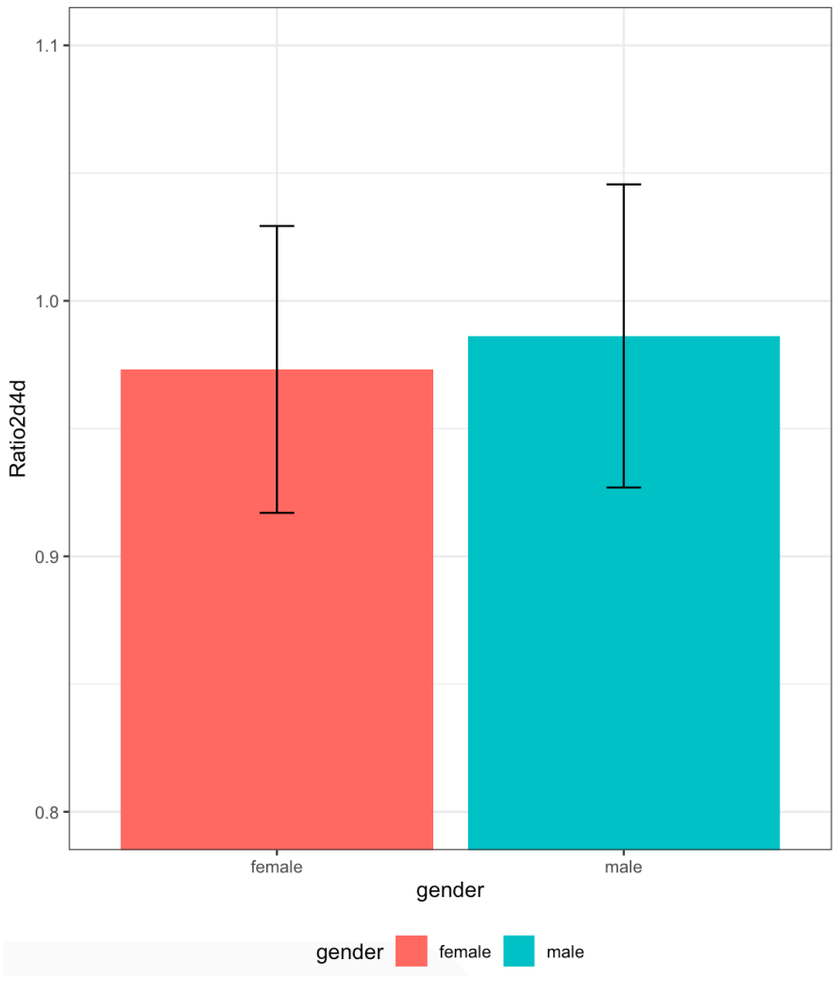
- Often we only have one sample instead of a sampling distribution of means

- If n > 30 and distribution ≈ normal:

- $$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}} \approx \frac{s}{\sqrt{n}}$$

Error bars: A way of representing variability

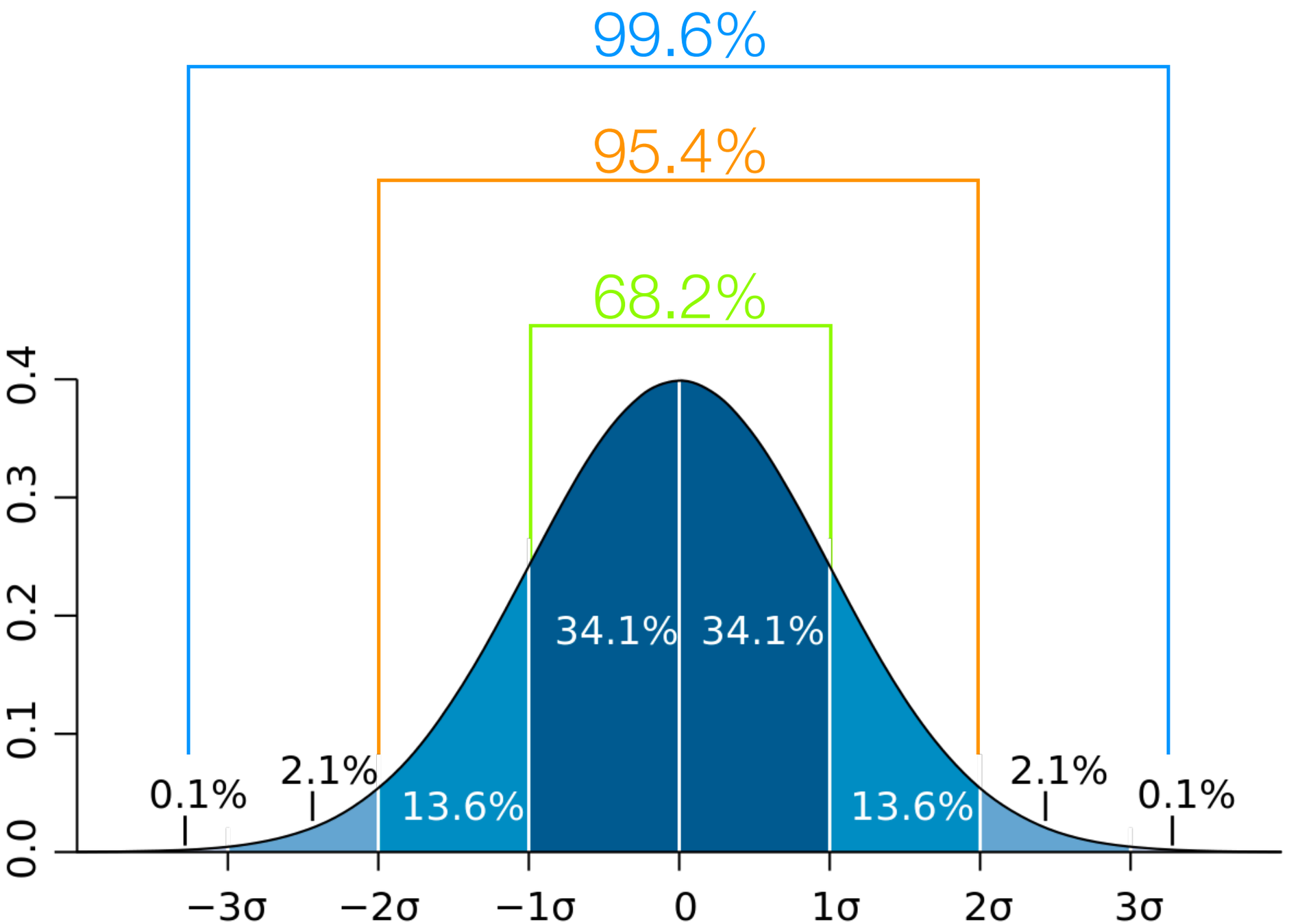
Standard Deviation



Measure of variability in the sample

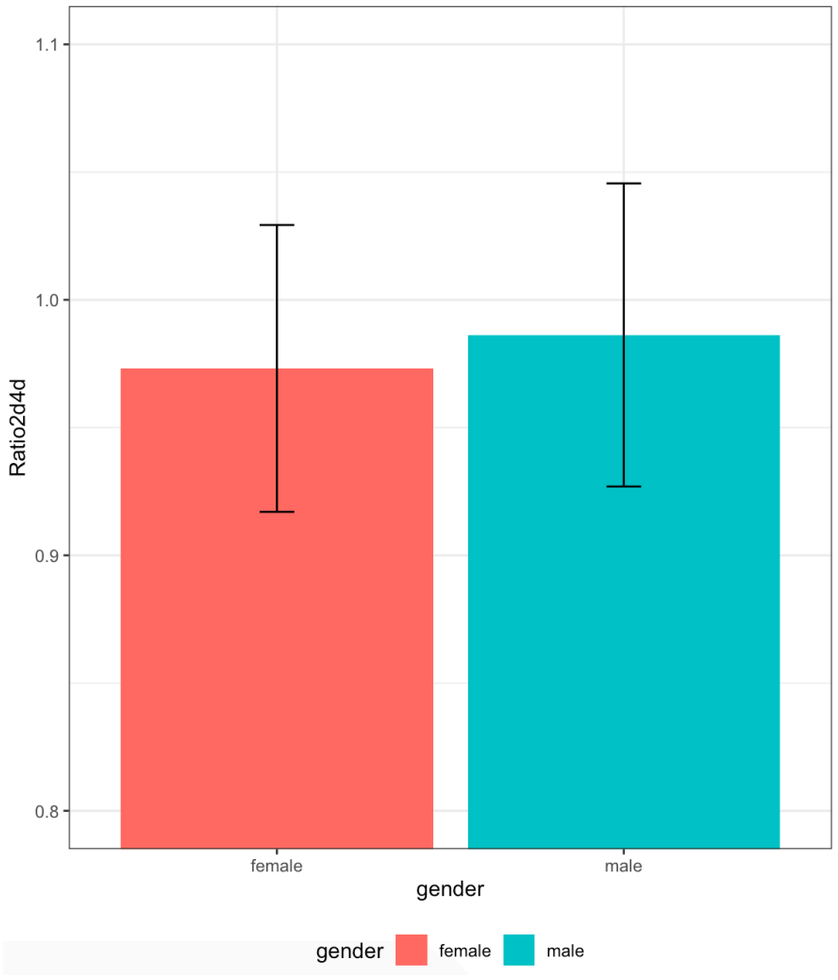
- Quantify the variability among the values
- Looking at whether the error bars overlap lets you compare the difference between the mean with the amount of scatter within the groups
- Knowing whether SD error bars overlap or not does not let you conclude whether the difference between the means is statistically significant or not
- **Descriptive, not inferential**

Interpreting *s*



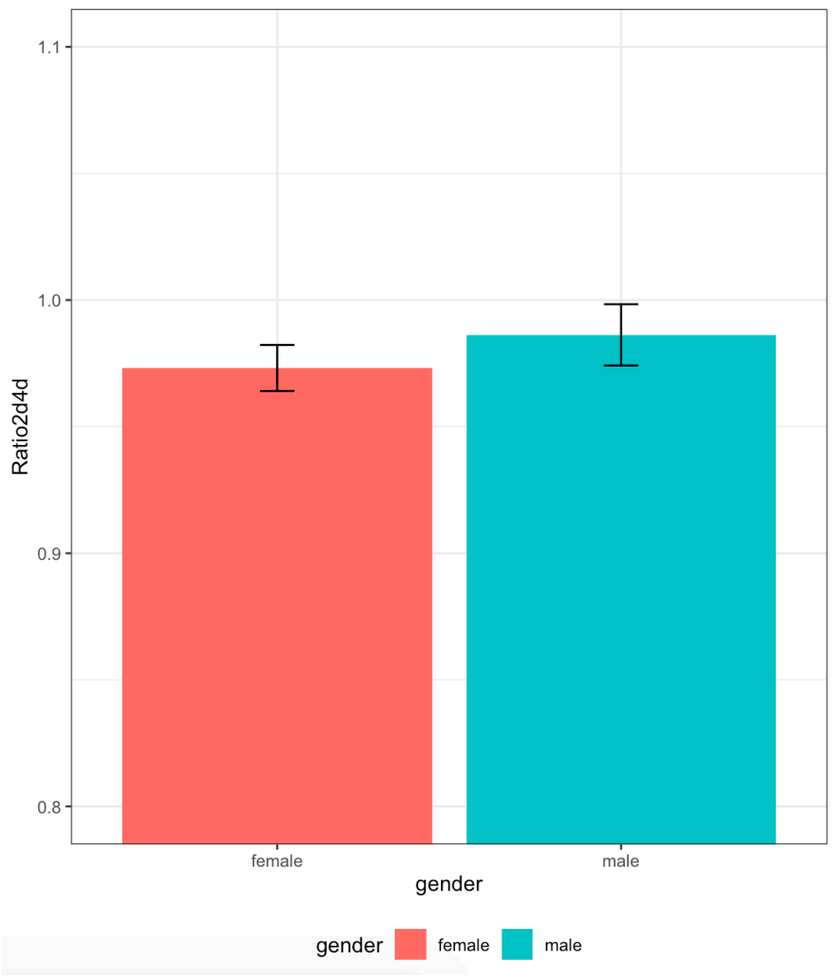
Error bars: A way of representing variability

Standard Deviation



Measure of variability in the sample

Standard Error (of the mean)



Measure of how well my sample mean approximates the true population mean

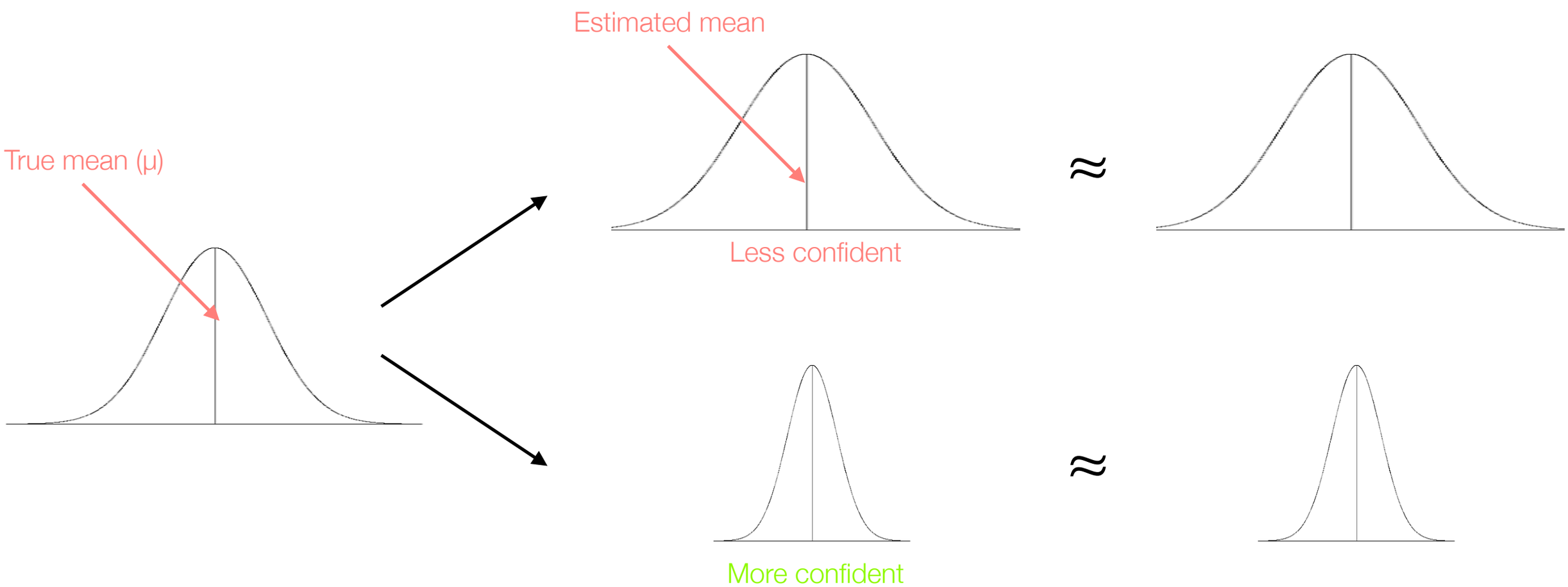
- Quantify how precisely you know the “true mean” of the population
- Looking at whether the error bars overlap lets you compare the difference between the mean with the precision of those means
- Large error bars mean that my sample mean is not a good estimate of the true population mean (= a lot of noise in the data)
- **Inferential, not descriptive**

Interpreting *SEM*

Population distribution

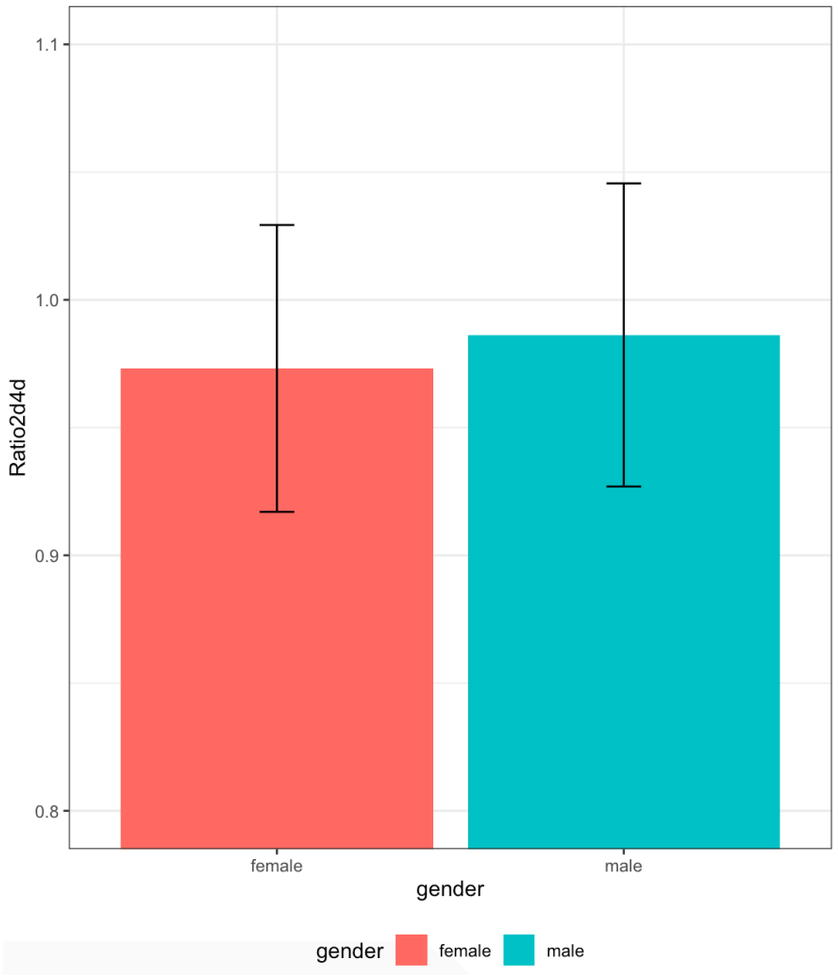
Theoretical distribution of many sample means

Distribution of observed sample



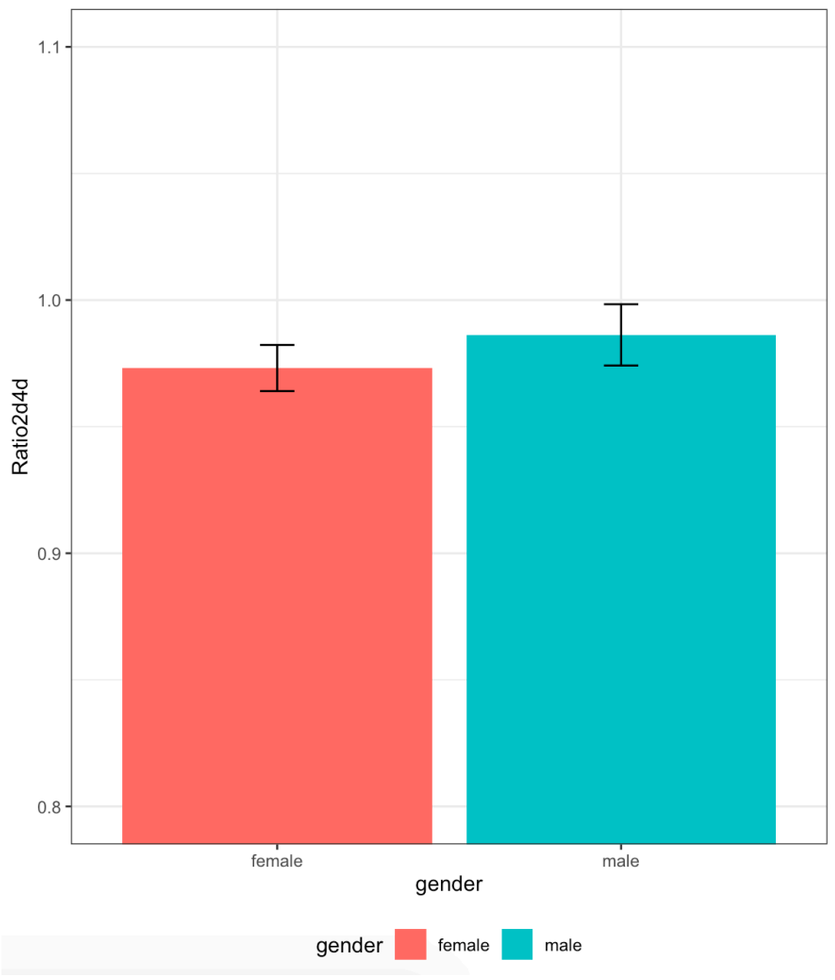
Error bars: A way of representing variability

Standard Deviation



Measure of variability in the sample

Standard Error (of the mean)



Measure of how well my sample mean approximates the true population mean

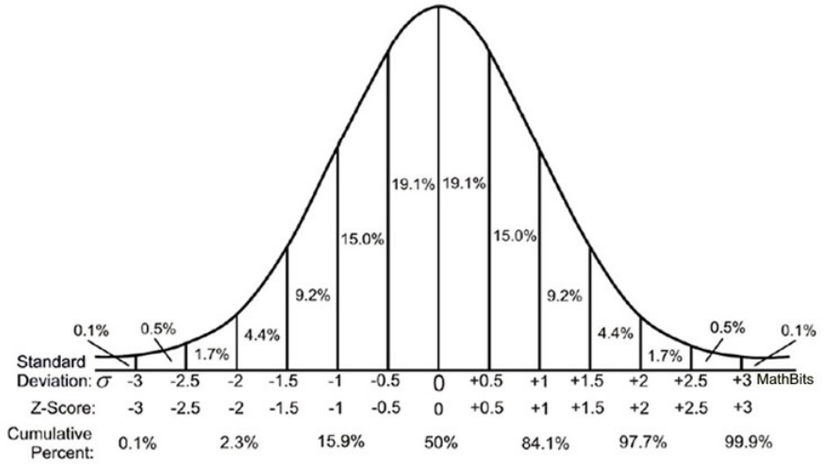
Confidence Intervals

Normalizing data: the Standard Normal Distribution

- = Expressing a score in terms of how many standard
  deviations it lays away from the mean

- → *z* scores

- Why?



- To compare across variables, e.g. when one is normally distributed and the other isn't

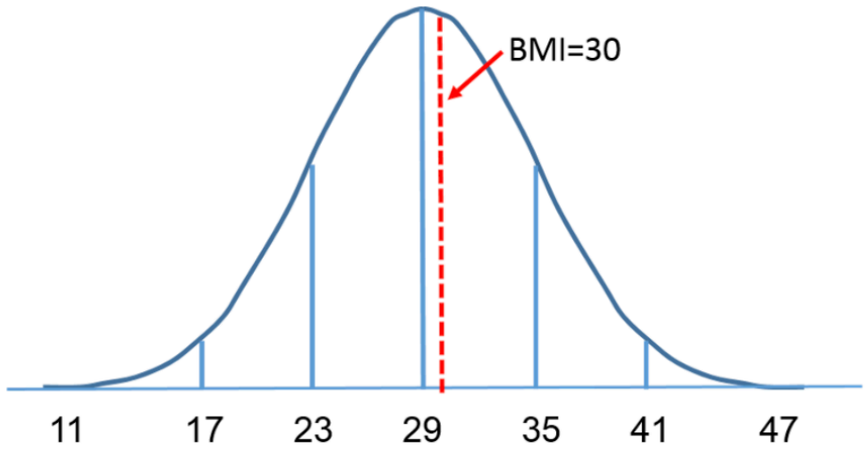
- To identify outliers

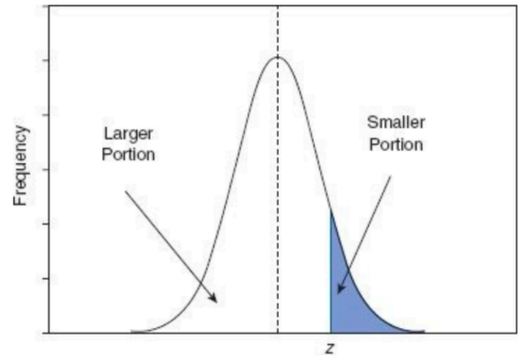
- To report results whenever we only care about the distribution of scores rather than the actual mean

- To use our data predictively (probability density function)

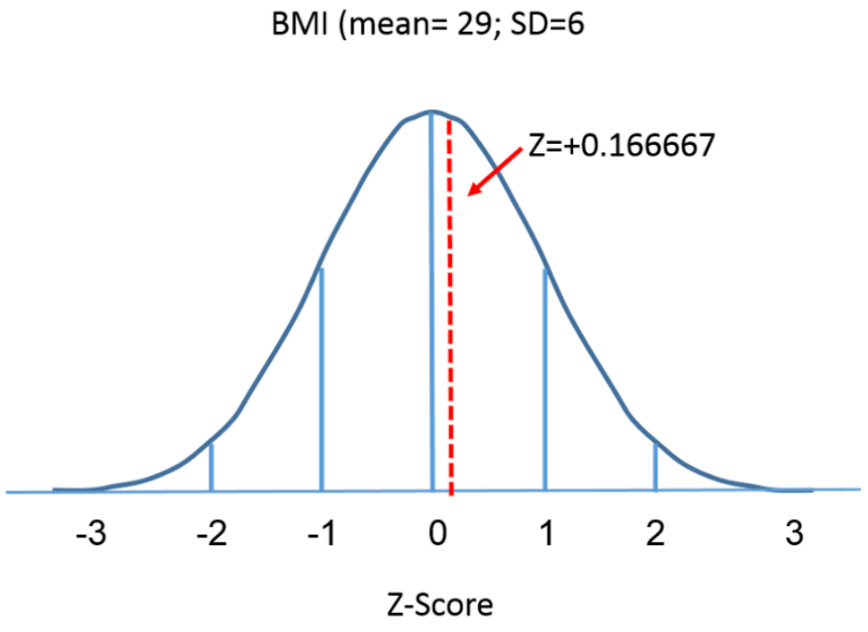
The z-score transformation (1)

- Standard Normal Distribution = probability distribution of z scores
- $\mu = 0, \sigma = 1$
- Probability distribution: area under the curve is 100%
- What's the probability of getting a certain score?
- $z = \dfrac{X - \bar{X}}{s}$
- $z \rightarrow P$





z	Large	Small	y
.12	.54776	.45224	.3961
.13	.55172	.44828	.3956
.14	.55567	.44433	.3951
.15	.55962	.44038	.3945
.16	.56356	.43644	.3939
.17	.56749	.43251	.3932
.18	.57142	.42858	.3925
.19	.57535	.42465	.3918
.20	.57926	.42074	.3910
.21	.58317	.41683	.3902
.22	.58706	.41294	.3894
.23	.59095	.40905	.3885
.24	.59483	.40517	.3876
.25	.59871	.40129	.3867
.26	.60257	.39743	.3857
.27	.60642	.39358	.3847

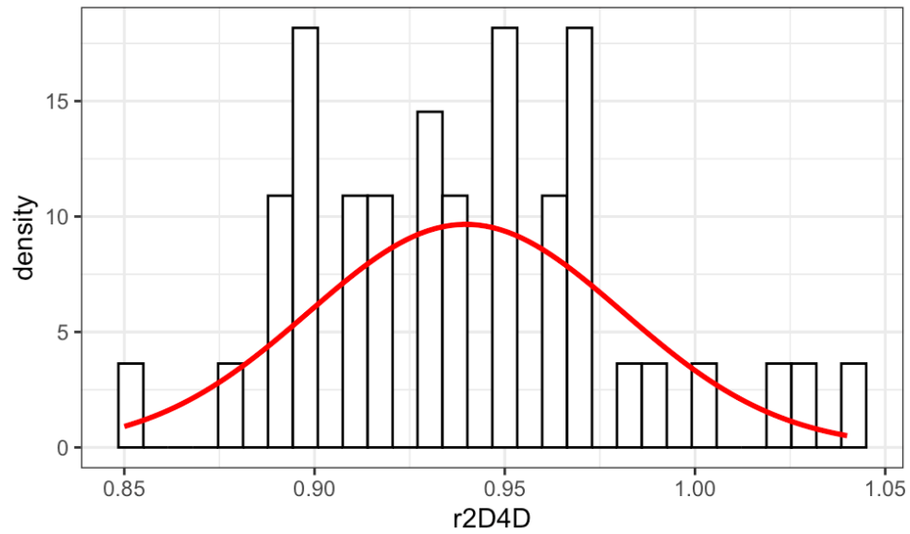


Does this look familiar?

and this?

The z-score transformation (2)

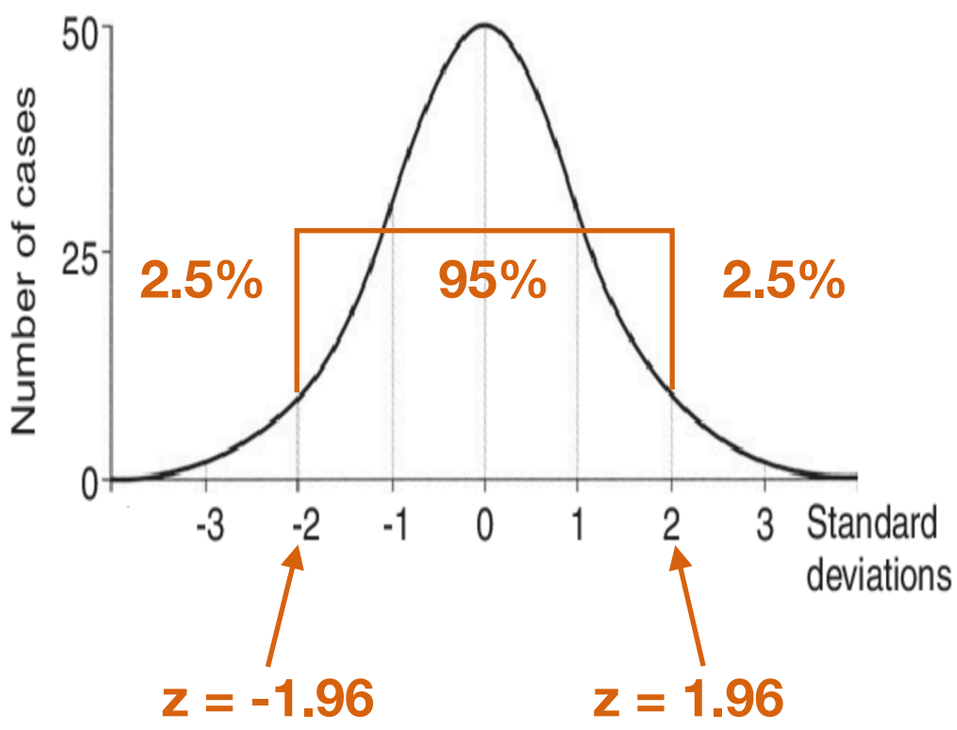
- What's the probability of having a 2D: 4D ratio of 1.13?
- $$z = \frac{X - \bar{X}}{s} = \frac{1.13 - 0.978}{0.057} = 2.65$$
- *P*(2D:4D ≧ 0.4%)
- *P*(2D:4D ≦ 99.6%)

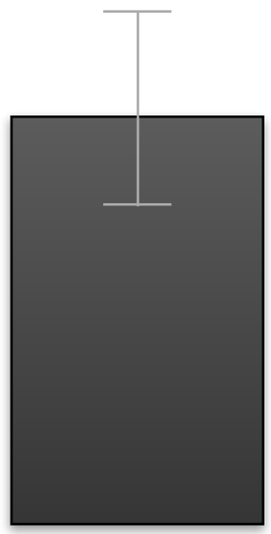


2.60	.99534	.00466	.0136
2.61	.99547	.00453	.0132
2.62	.99560	.00440	.0129
2.63	.99573	.00427	.0126
2.64	.99585	.00415	.0122
2.65	.99598	.00402	.0119
2.66	.99609	.00391	.0116

Confidence intervals

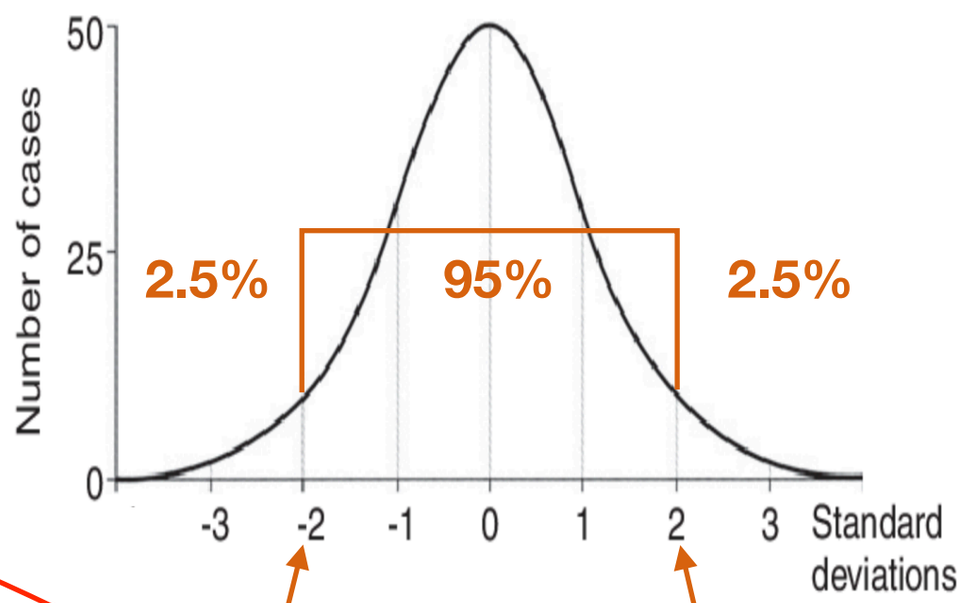
- A range of values within which we estimate the true value of the mean to fall
- We estimate that in 95% (99%) of samples, the true mean population mean will fall within this range
- Formula: $\bar{x} \pm z \times \frac{\sigma}{\sqrt{n}}$
  - $upper = \bar{x} + (1.96 \times SE)$
  - $lower = \bar{x} - (1.96 \times SE)$
- Basically: a measure of the accuracy of my sample mean (SEM) within 2SD around the sample mean
- If $\mu_{\bar{x}} \approx \bar{x}$, then confidence intervals will be small

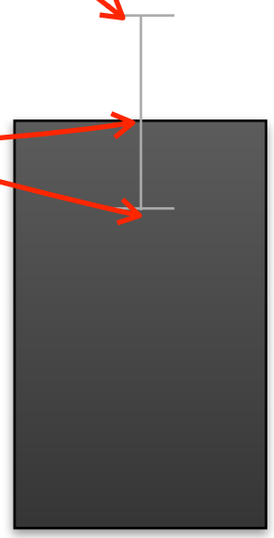




Confidence intervals

- A range of values within which we estimate the true value of the mean to fall
- We estimate that in 95% (99%) of samples, the true mean population mean will fall within this range
- Formula: $\bar{x} \pm z \times \frac{\sigma}{\sqrt{n}}$
  - $upper = \bar{x} + (1.96 \times SE)$
  - $lower = \bar{x} - (1.96 \times SE)$
- Basically: a measure of the accuracy of my sample mean (SEM) within 2SD around the sample mean
- If $\mu_{\bar{x}} \approx \bar{x}$, then confidence intervals will be small



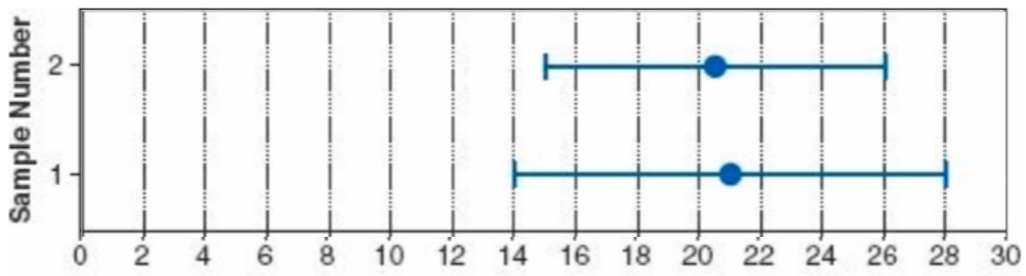


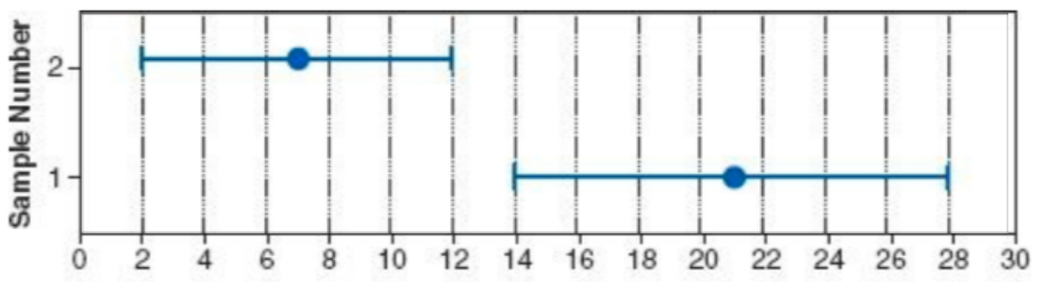
Comparing different means

- Non-overlapping error bars suggest that the samples come from different populations

- If our experimental manipulation is successful, we expect our samples to come from different populations

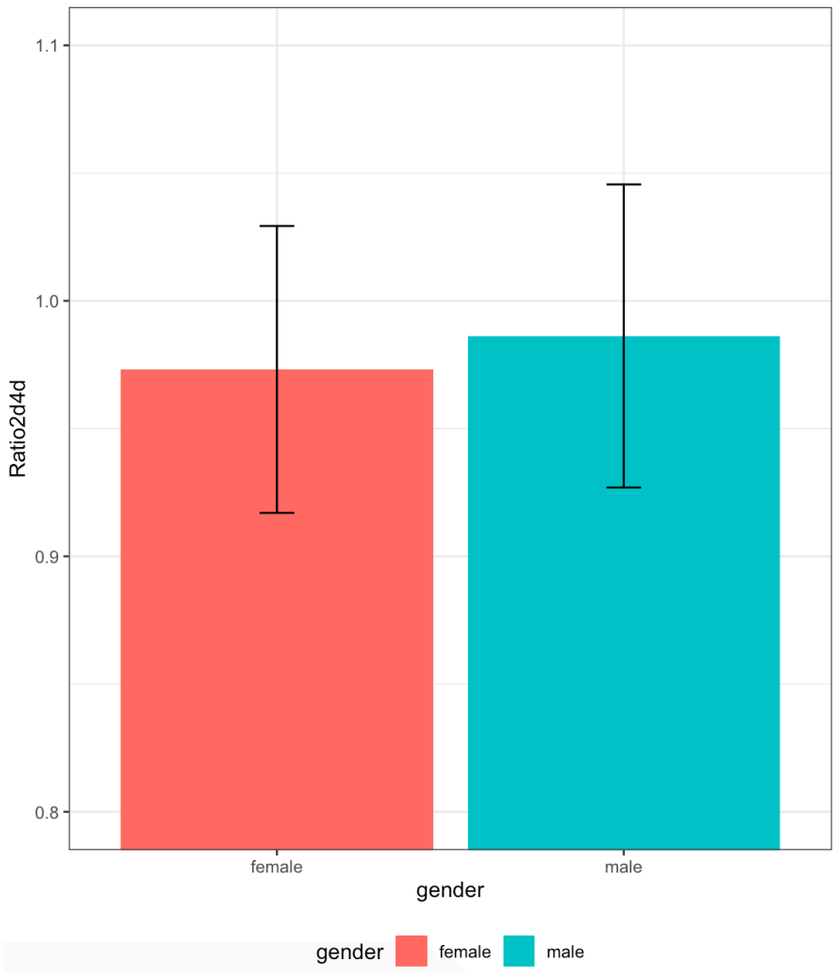
- “True effect” vs. “spurious effect”





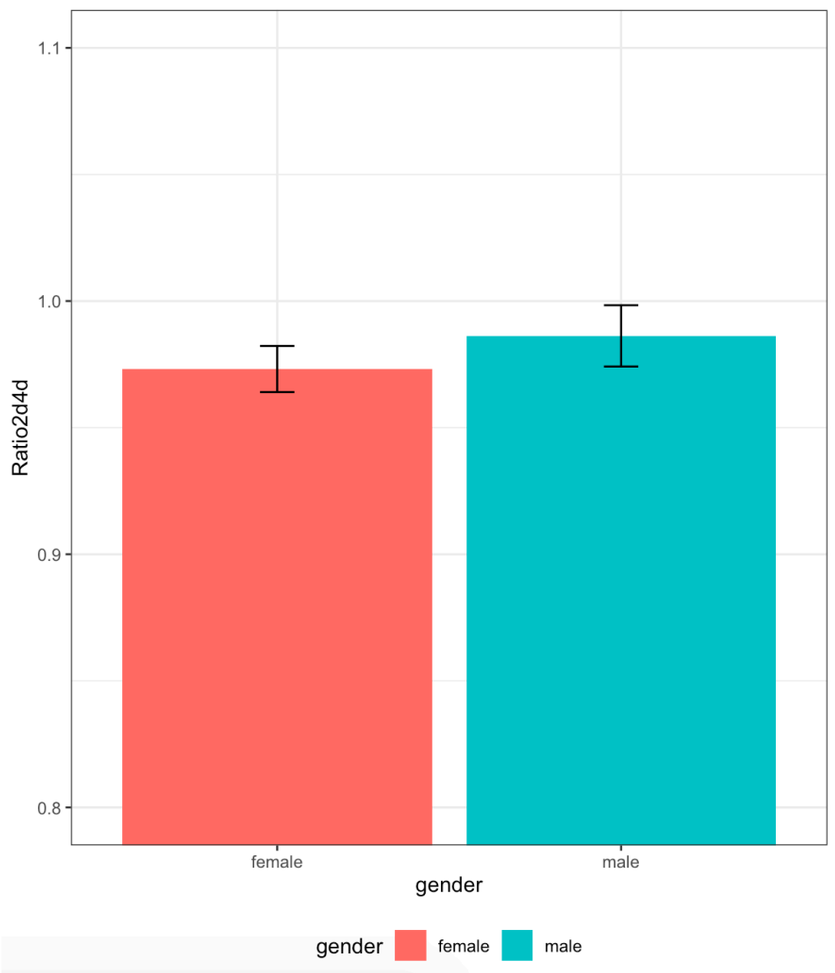
Error bars: A way of representing variability

Standard Deviation



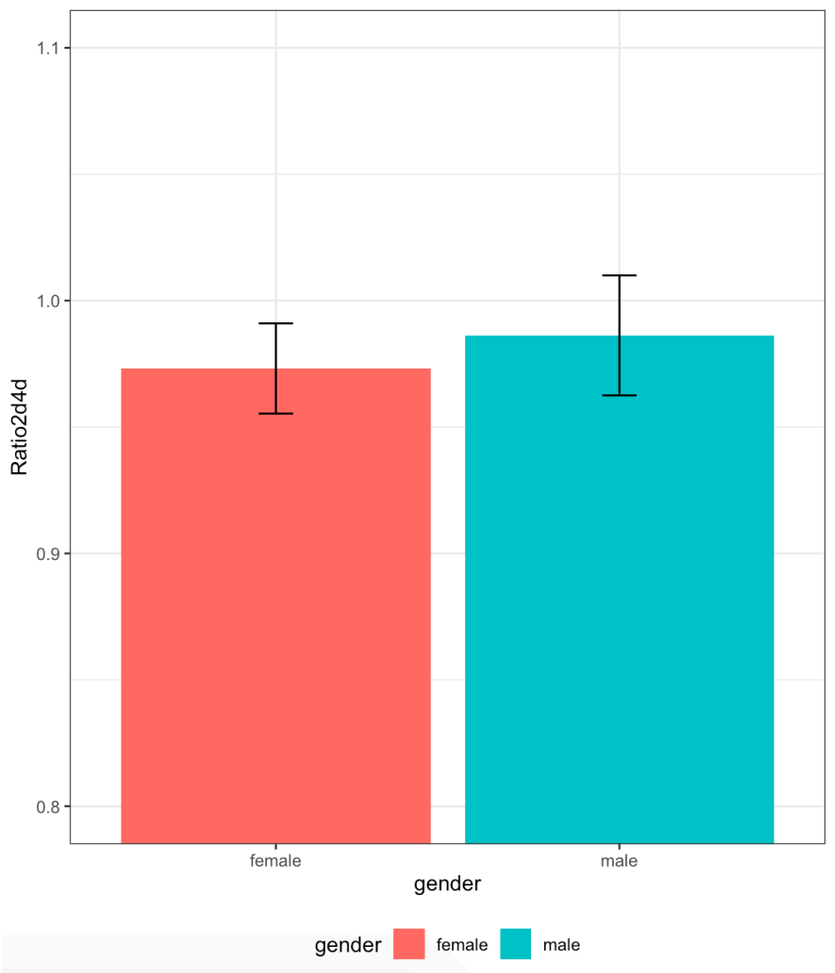
Measure of variability in the sample

Standard Error (of the mean)



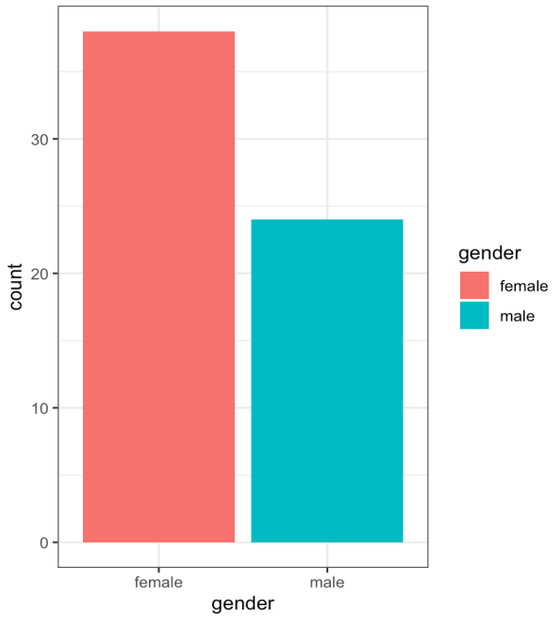
Measure of how well my sample mean approximates the true population mean

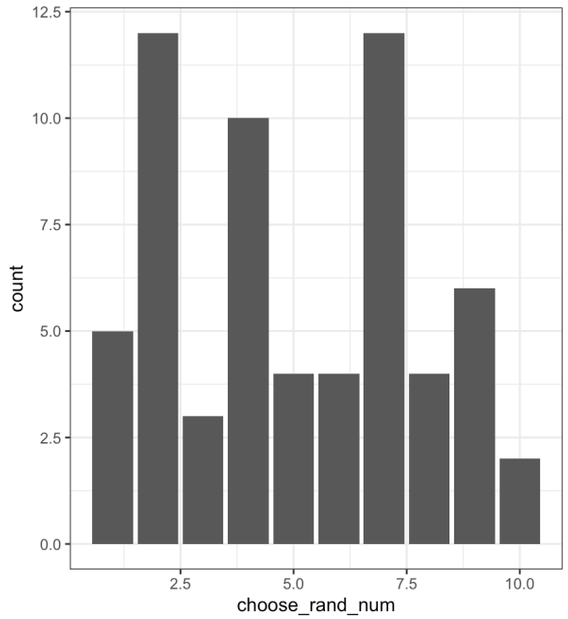
Confidence Intervals

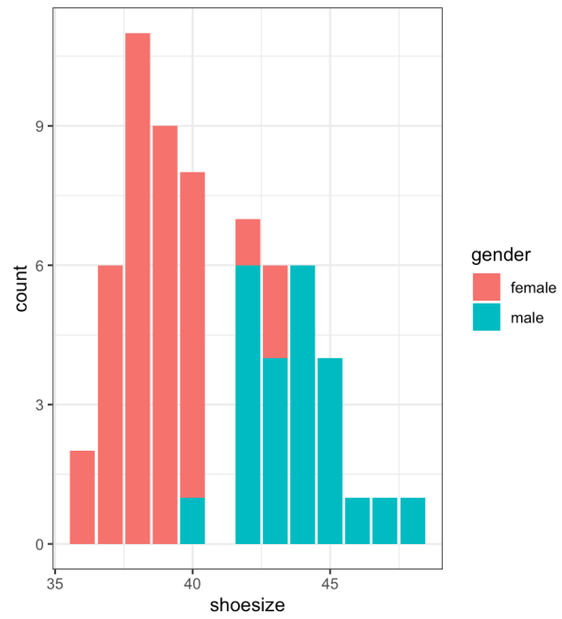


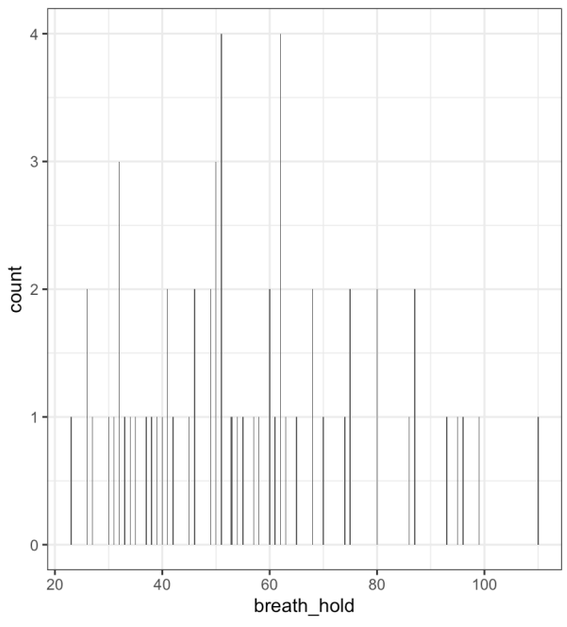
Boundaries within which we believe the true mean of the population to fall some % of the time (usually 95%)

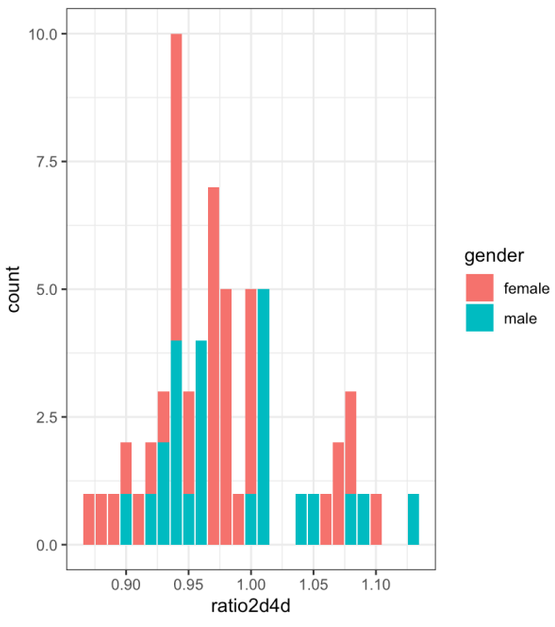
Recap: Histogram as frequency distribution

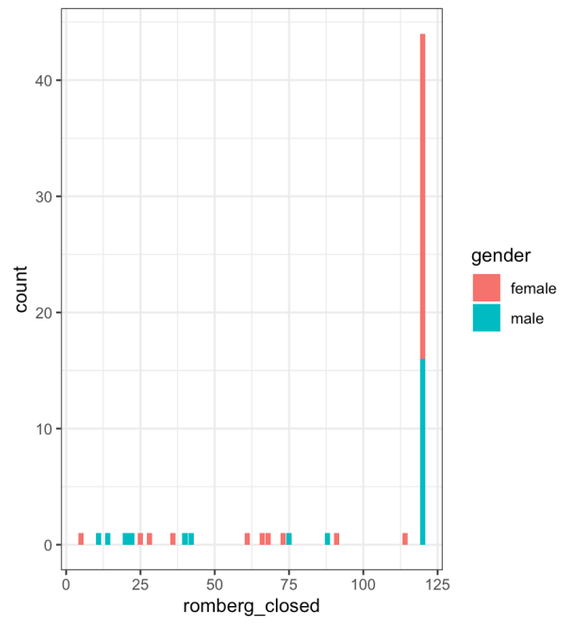






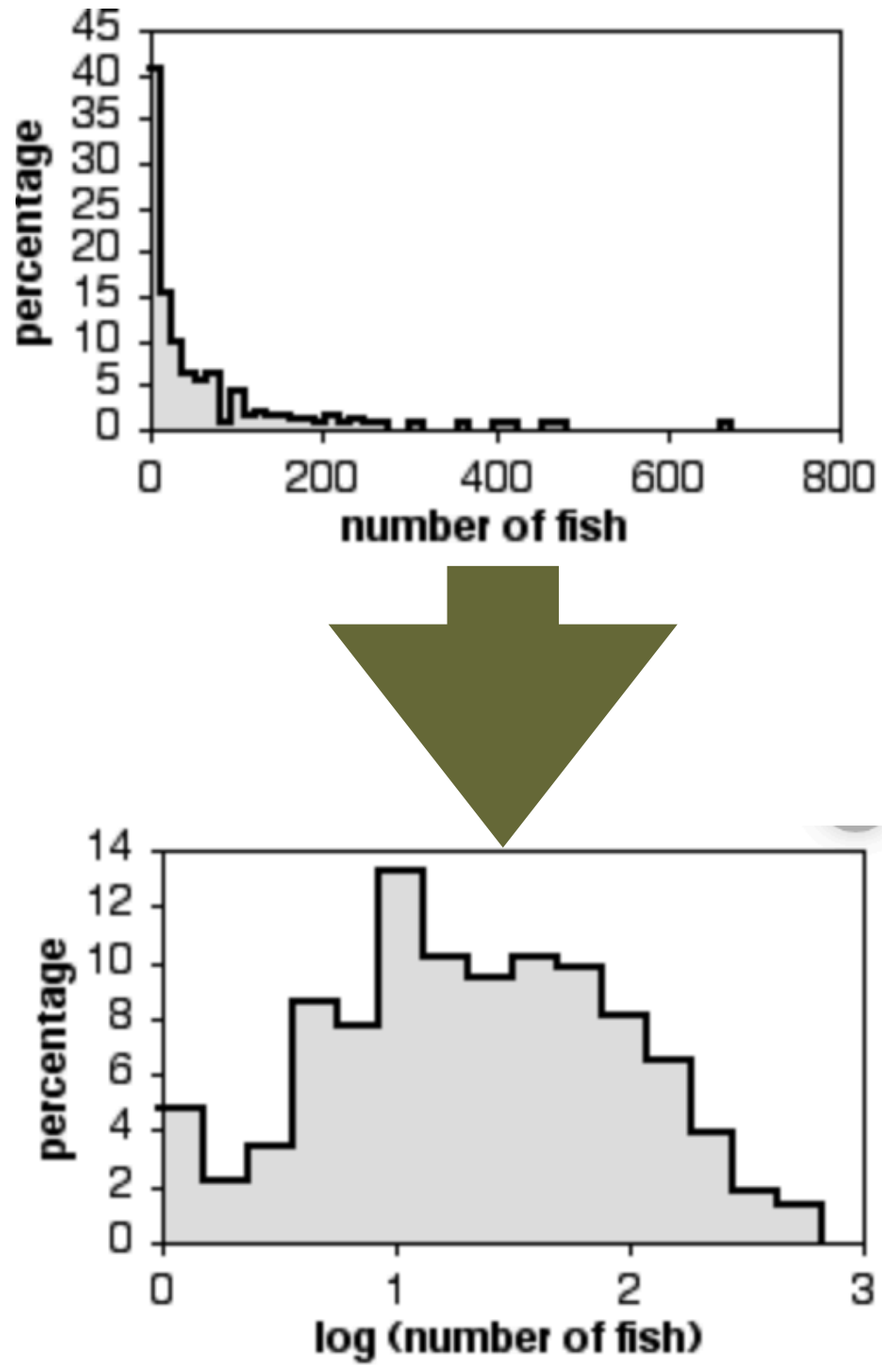






Assumptions of normality in statistical texts

- **Parametric tests**: assume that data comes form population with fixed parameters, eg:
  - *t*-test, correlation, ANOVA
    - these won't work with non-normal data
- **Non-parametric tests**: "distribution-free"
- Data can be *transformed* to become normally distributed



Assumptions of parametric tests

1. Data are **normally distributed**

2. **Variance is homogeneous** across samples, groups, levels of a variable

3. Data are at least at the **interval level**

4. **Data are independent** from each other across participants or across sessions within participants

Bonus slide: Variable types

- Categorical
  - Binary/Logical (frequency)
  - Nominal (frequency)
  - Ordinal (frequency + order)
- Continuous
  - Interval (full arithmetic)
  - Ratio (full arithmetic)

POS	NO	DRIVER	CAR
1	5	Sebastian Vettel	FERRARI
2	16	Charles Leclerc	FERRARI
3	33	Max Verstappen	RED BULL RACING HONDA
4	44	Lewis Hamilton	MERCEDES
5	77	Valtteri Bottas	MERCEDES
6	23	Alexander Albon	RED BULL RACING HONDA
7	4	Lando Norris	MCLAREN RENAULT
8	10	Pierre Gasly	SCUDERIA TORO ROSSO H

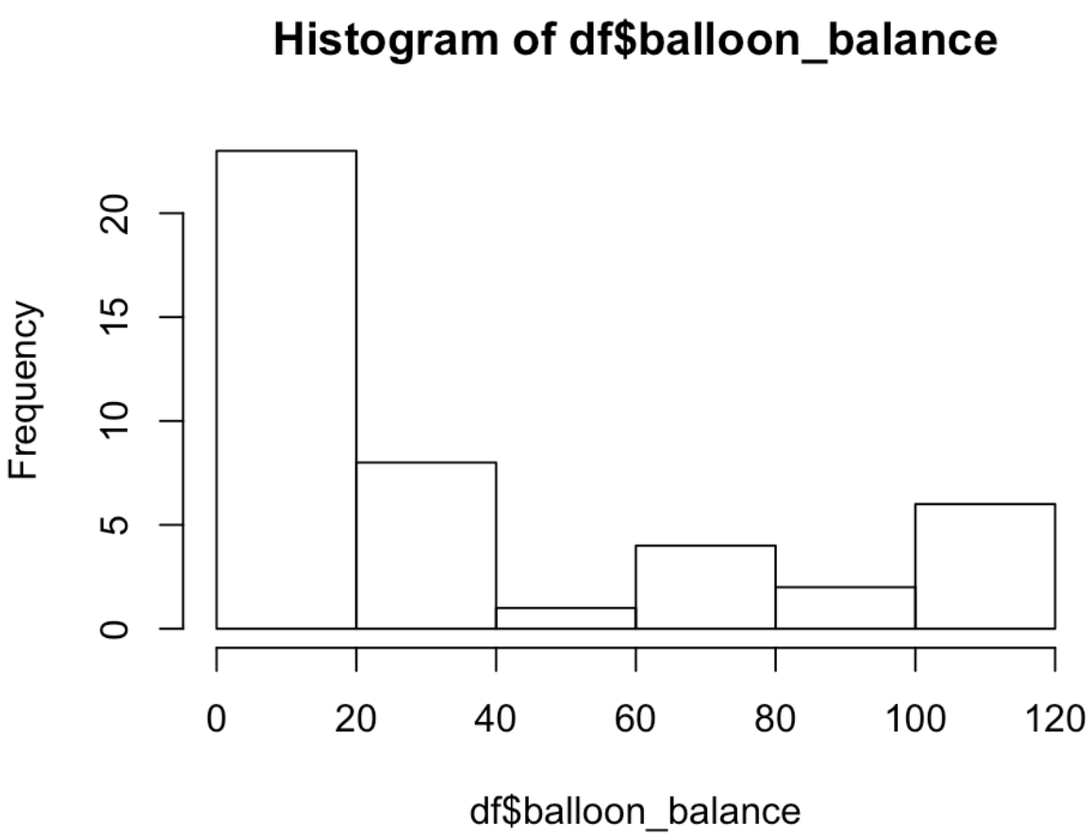
Eg. Parametric vs non-parametric group mean comparisons

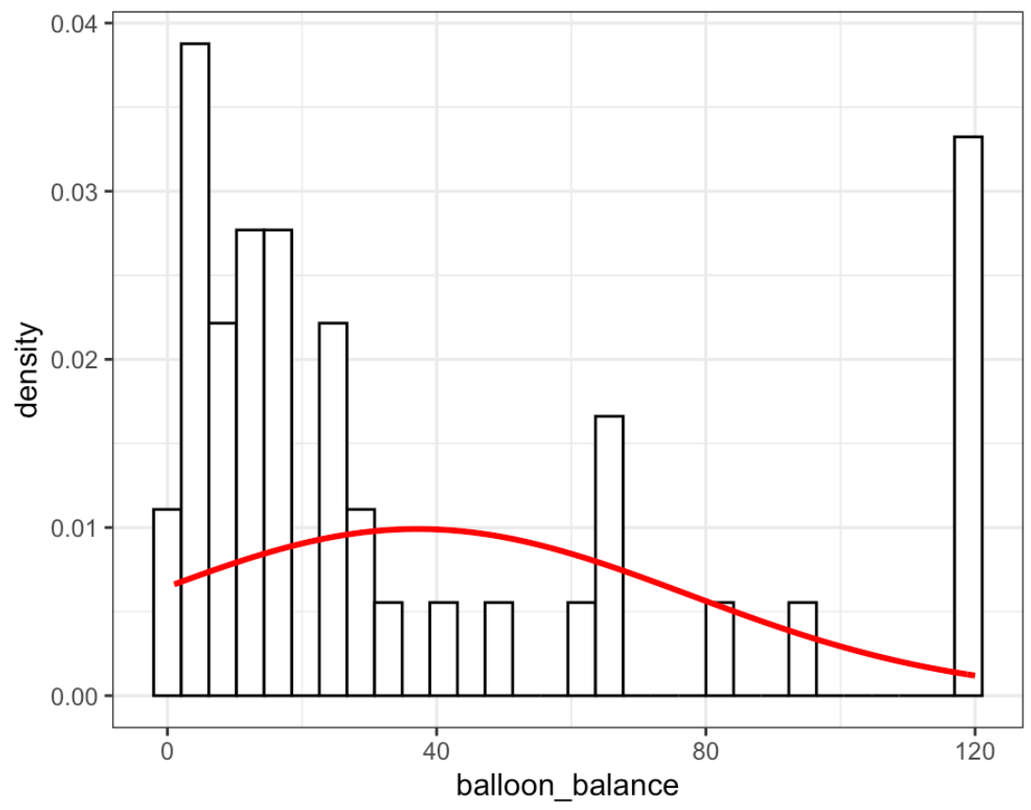
Parametric test	Non-Parametric equivalent
Paired t-test	Wilcoxon Rank sum Test
Unpaired t-test	Mann-Whitney U test
Pearson correlation	Spearman correlation
One way Analysis of variance	Kruskal Wallis Test

Variable	Test
Nominal	McNemar's Test
Ordinal (Ordered categories)	Wilcoxon
Quantitative (Discrete or Non-Normal)	Wilcoxon
Quantitative (Normal*)	Paired *t* test

Assessing normality through visual inspection (1)

- Histogram



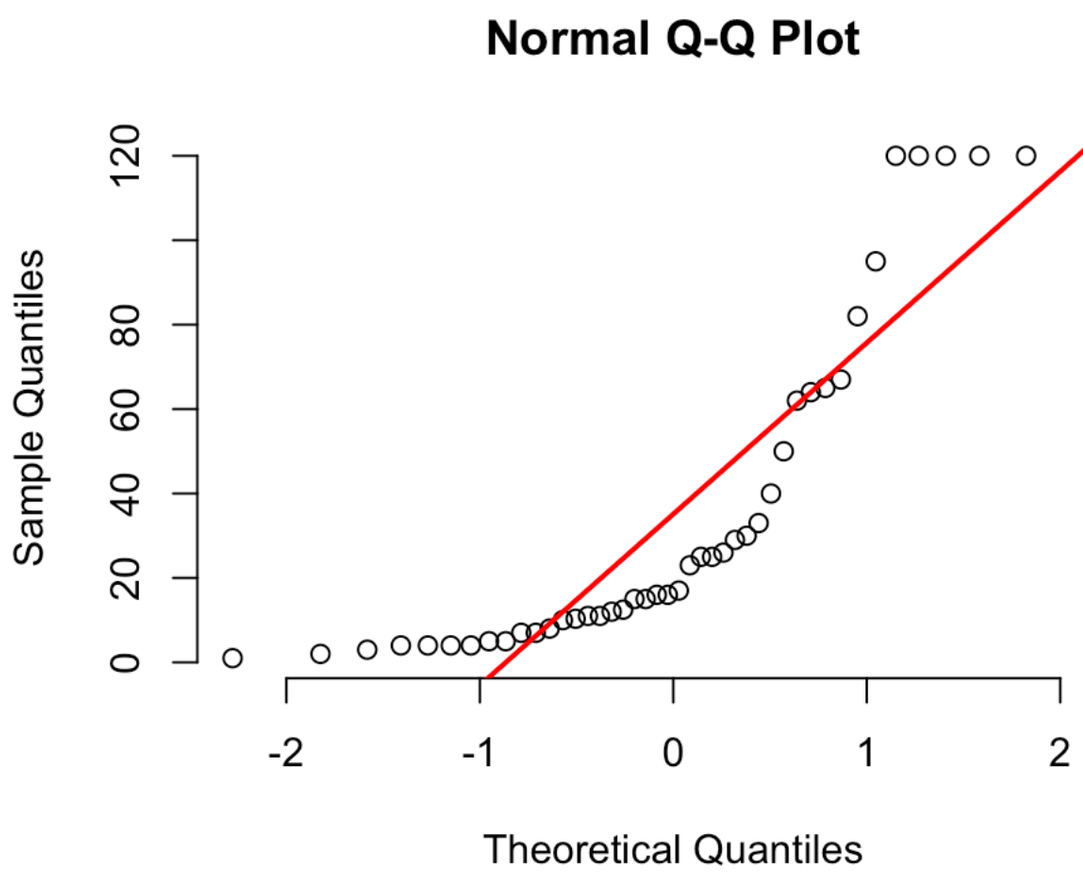



```
hist(df$balloon_balance)
```

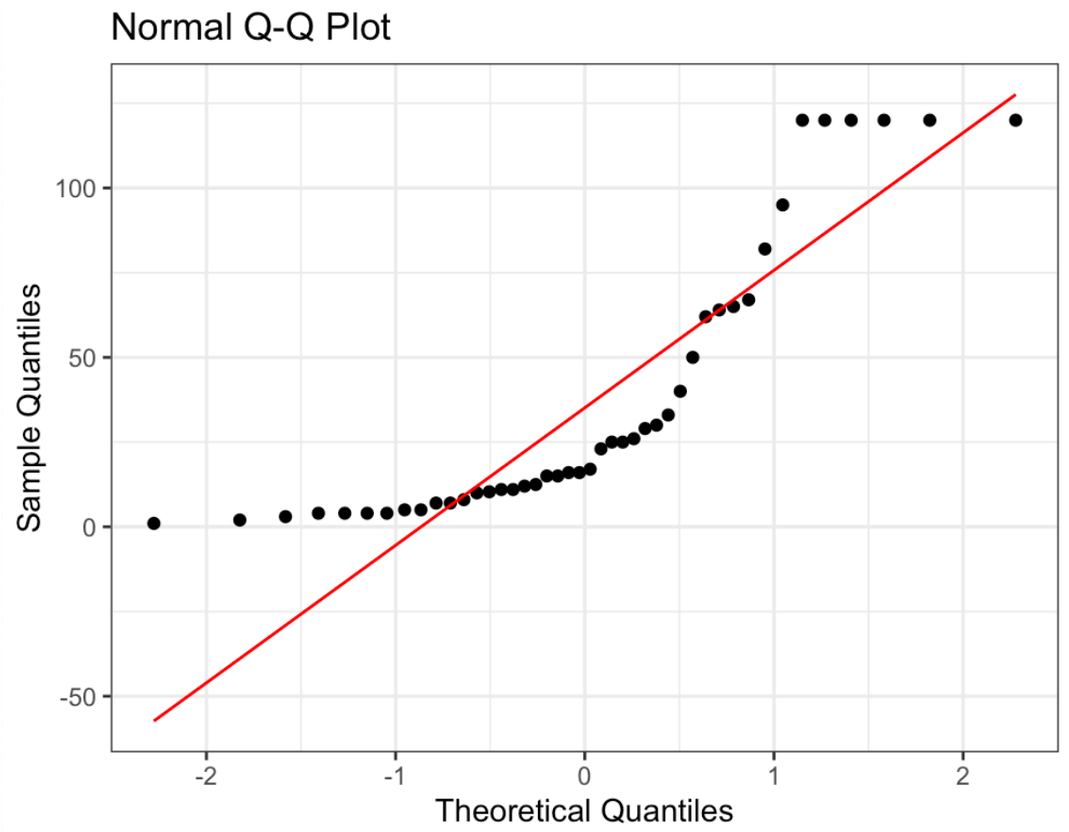
```
ggplot(df, aes(balloon_balance)) +
      geom_histogram(aes(y = ..density..), colour = "black", fill = "white") +
      stat_function(fun = dnorm, args = list(mean = mean(df$balloon_balance),
            sd = sd(df$balloon_balance)), colour = "red", size = 1) +
      theme_bw()
```

Assessing normality through visual inspection (2)

- Quantile-Quantile plot (QQ-plot)



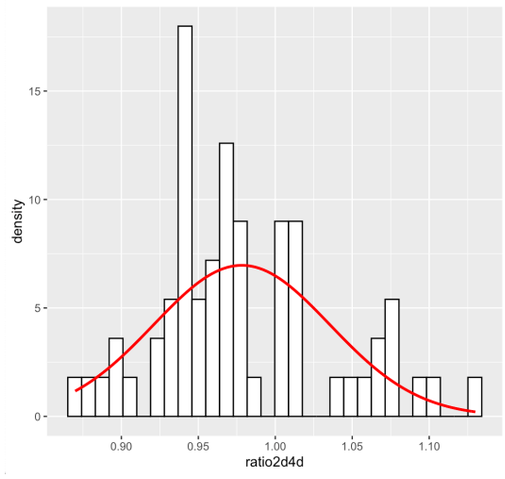

```
qqnorm(df$balloon_balance, pch = 1, frame = FALSE)
qqline(df$balloon_balance, col = "red", lwd = 2)
```

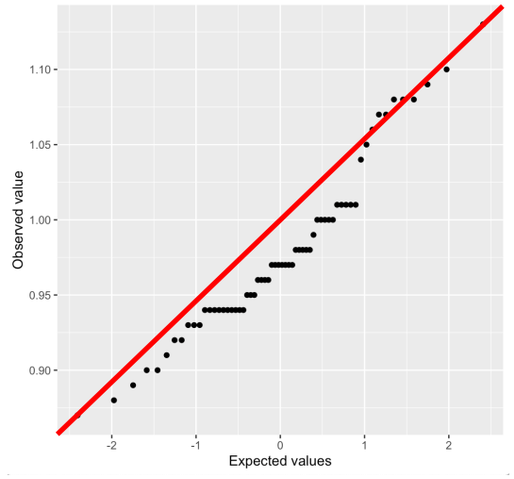



```
ggplot(df, aes(sample = balloon_balance)) +
    stat_qq() +
    stat_qq_line(colour = "red") +
    labs(x = "Theoretical Quantiles", y = "Sample Quantiles") +
    ggtitle("Normal Q-Q Plot") +
    theme_bw()
```

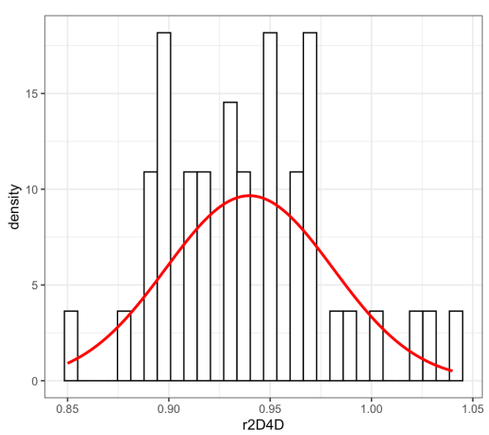
Example of close-to-normal distributions (1)

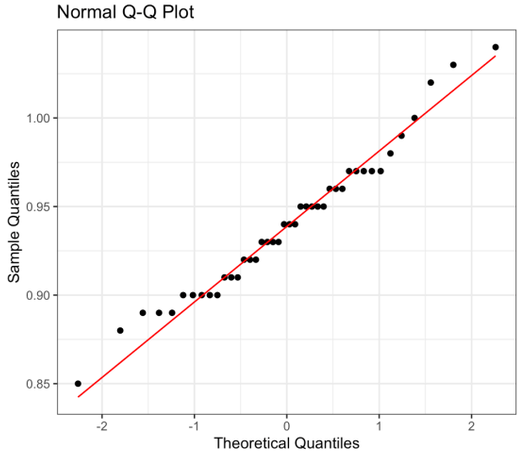
CogSci2019



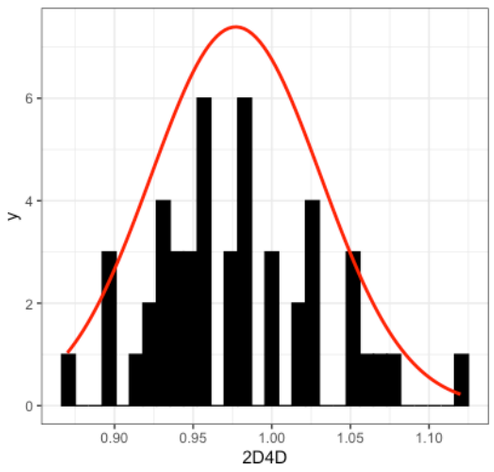


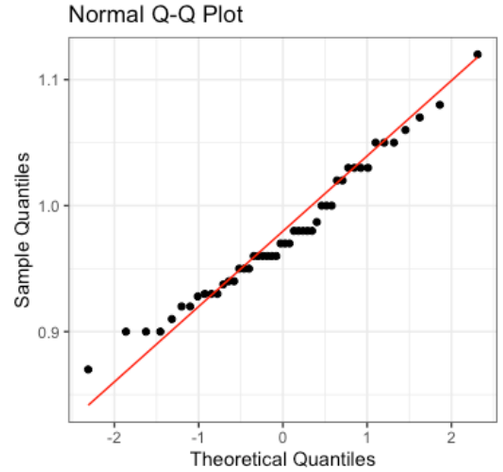
CogSci2020



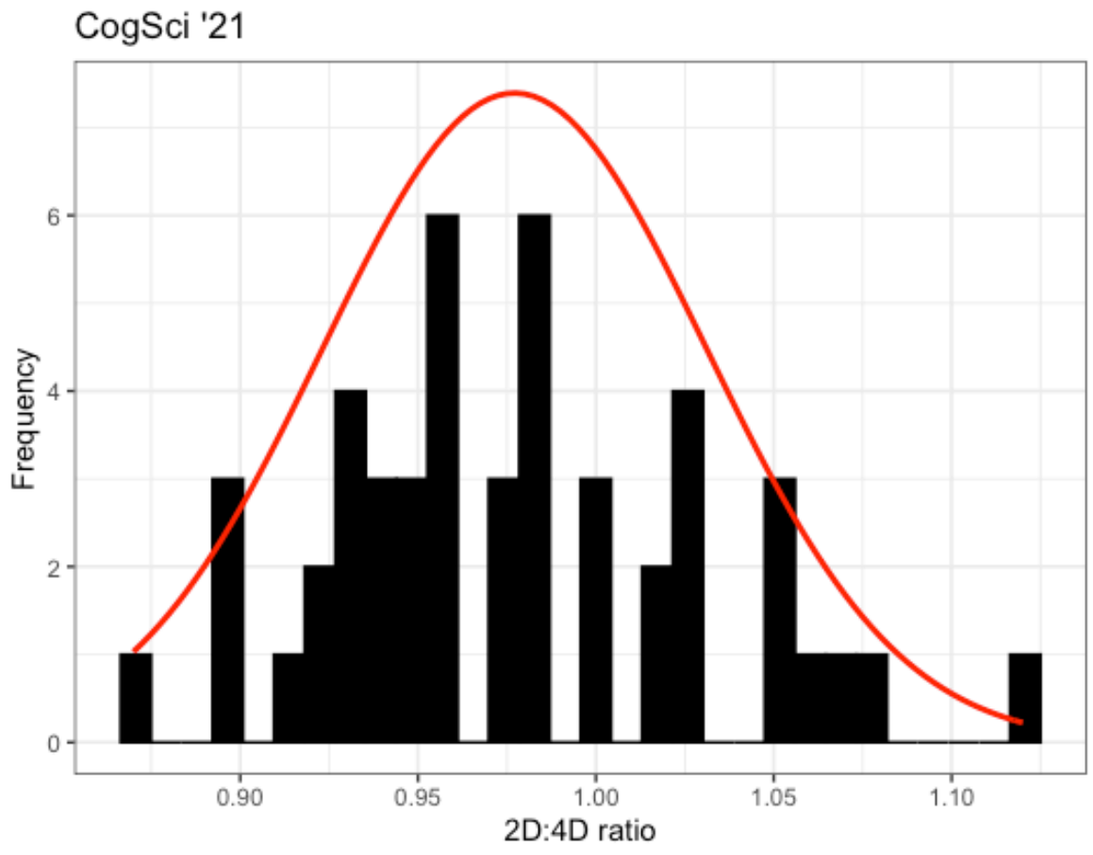


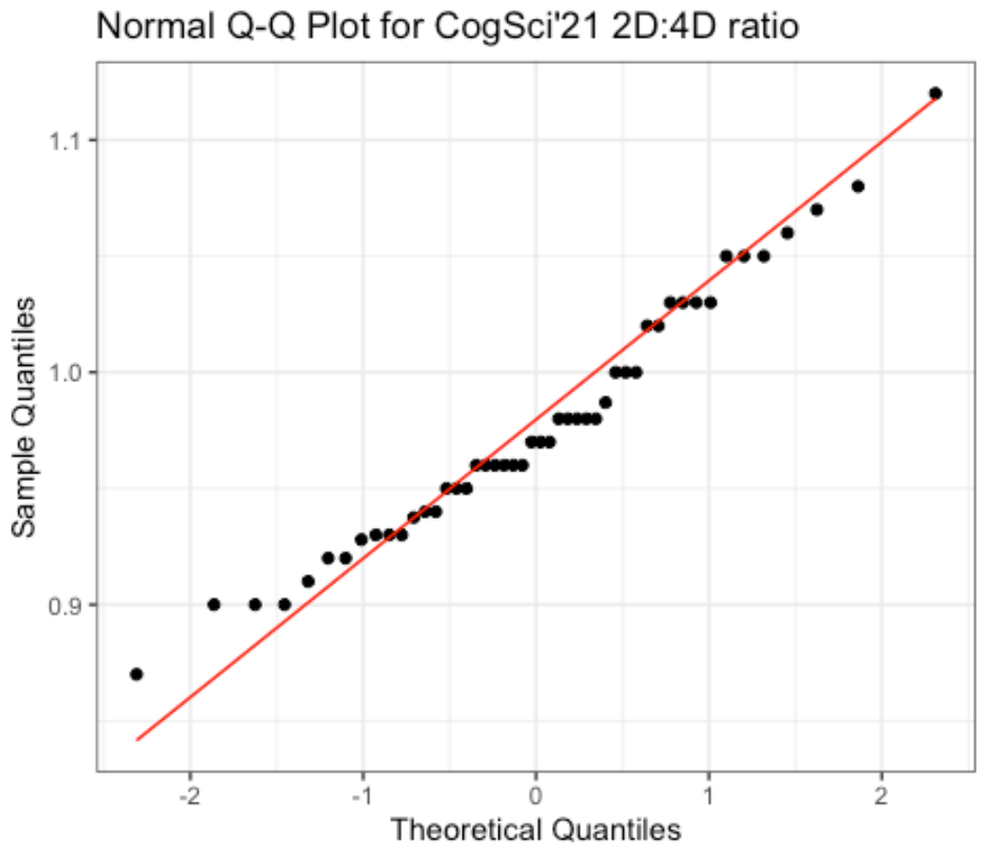
CogSci2021





Example of close-to-normal distributions (2)





Assessing data normality

- Statistical test:
  - pastecs::stat.desc()
- Skewness/kurtosis
  - If skew.2SE and kurt.2SE > 1
    → no normality
- Shapiro-Wilk test of normality
  - if $p < 0.05$ → significantly different from normal

```
round(pastecs::stat.desc(cbind(df$b
alloon_balance, df$r2d4d), basic =
FALSE, norm = TRUE), digits = 2)
```

	V1	V2
median	16.50	0.94
mean	37.18	0.94
SE.mean	6.07	0.01
CI.mean.0.95	12.24	0.01
var	1620.05	0.00
std.dev	40.25	0.04
coef.var	1.08	0.05
skewness	1.11	0.40
skew.2SE	1.55	0.56
kurtosis	-0.25	-0.33
kurt.2SE	-0.18	-0.23
normtest.W	0.77	0.97
normtest.p	0.00	0.29

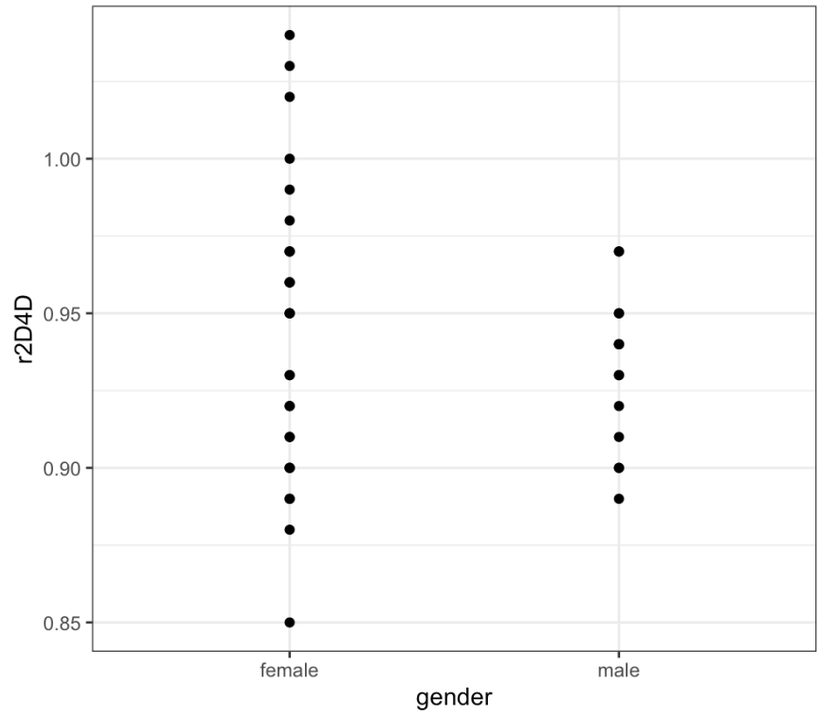
More on skewness/kurtosis:
https://bit.ly/33LhMhp

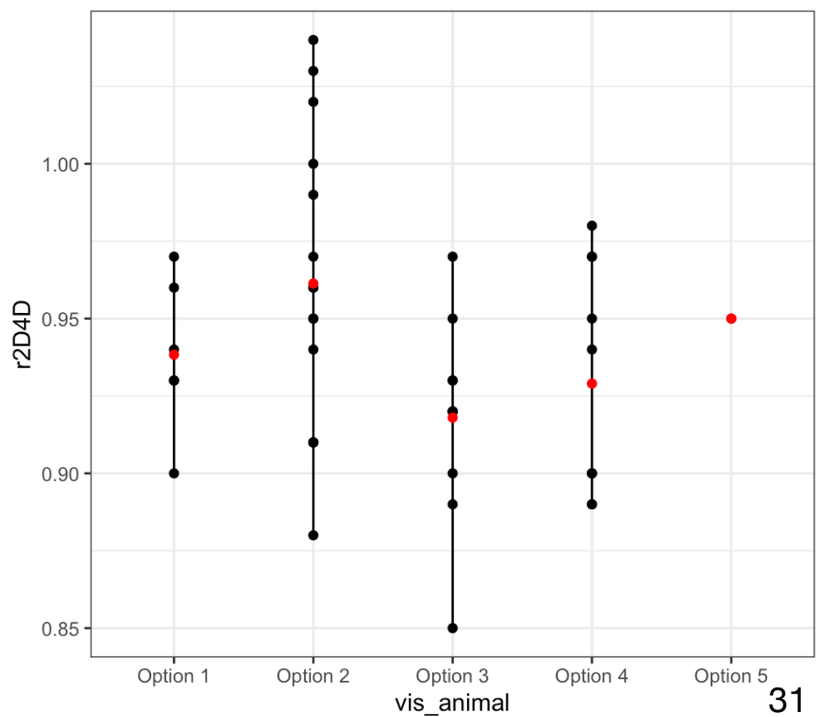
Assessing homogeneity/heterogeneity of variance

- Levene's test: is the H_1 of difference between variances statistically significant?

```
> car::leveneTest(df$r2D4D, df$gender, center = mean)
Levene's Test for Homogeneity of Variance (center = mean)
      Df F value  Pr(>F)
group  1  6.1773 0.01722 *
      40
```

```
> car::leveneTest(df$r2D4D, df$vis_animal, center = mean)
Levene's Test for Homogeneity of Variance (center = mean)
      Df F value Pr(>F)
group  4   1.668 0.1781
      37
```



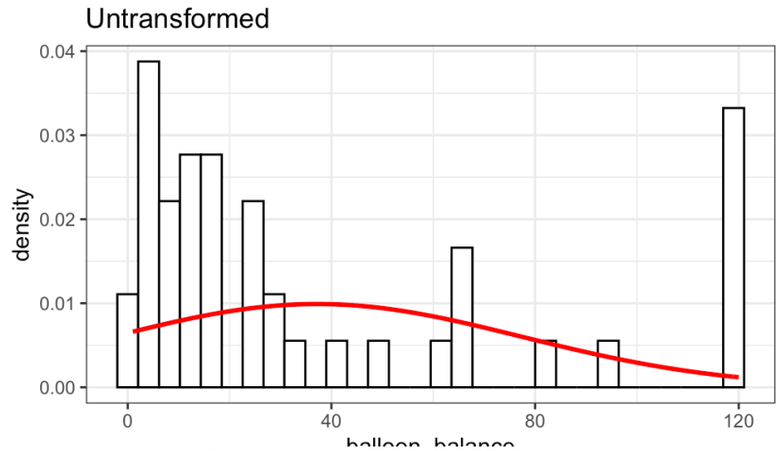


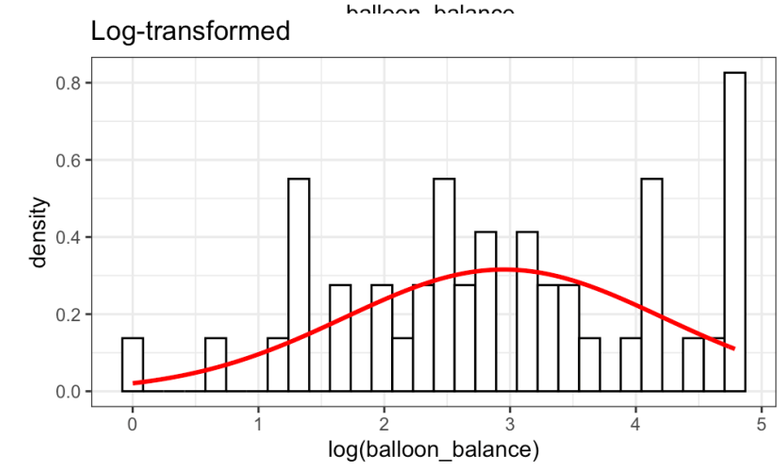
Common problems with the data

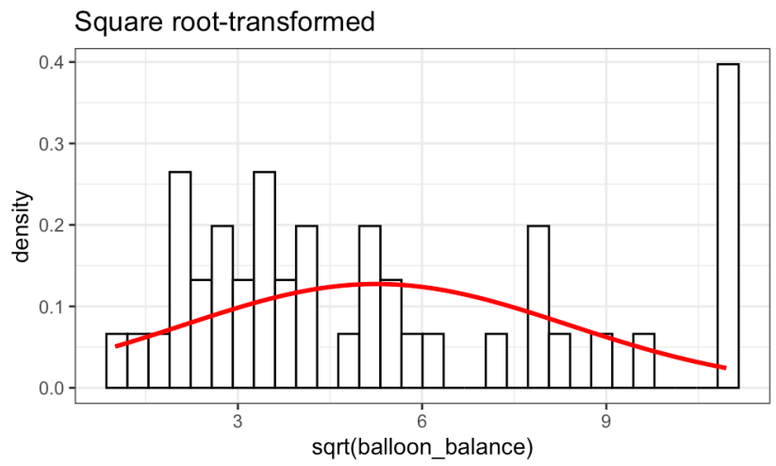
Cause	Description	Solution
Small sample size	Larger samples tend towards normality	Collect more data
Measurement error	Weird (impossible?) numbers in the dataset	Remove the odd one out / Rerun the experiment
Outliers	People performing especially well/ poorly	Remove outliers (e.g. > 3 SD)
Ceiling/floor effects	Task is too easy/hard	Throw everything in the trash and start again!
Skew/kurtosis/bimodality	Data may be generated by multiple interacting factors	Trasform the data / Choose statistical analyses that do not assume normality

Data transformation

Problem	Transformation	Limits	R syntax
- Positive skew - Unequal variances	log transformation: $\log(X_i)$	Can't deal with negative numbers	`log()`
- Positive skew - Unequal variances	square root transformation: $\sqrt{X_i}$	Bigger effect than log(), still can't deal with negative numbers	`sqrt()`
- Positive skew - Unequal variances	reciprocal transformation: $1/X_i$	Reduces large scores, good for negative numbers, but reverses scores	`1/x` `1/(max(x)-x)`
Negative skew	Reverse score transformations	Reverse the data ($\max(X)-X_i$) before running any of the above transformations	







Next week

- PsychoPy workshop
- Kristian Tylén will give you more info soon
- Make sure to download and install
  - Python 3
  - PsychoPy 3
- before showing up for the workshop

Also next week

- Portfolio 1 deadline:
- Data mining report on the personality test data
- Wednesday 29/9 at 23:59