01_statistical_inference - Week 3

Notes on the course
https://www.coursera.org/learn/statistical-inference

Andres FR

August 2020

T Confidence Intervals

Via CLT we have made confidence intervals of means in the form $\hat{\mu} \pm Z \cdot Q \cdot \hat{s}$ where Z is the normal dist. quantile, and $\hat{s} = \sqrt{\frac{Var(X)}{n}}$ is our estimated standard error.

When dealing with few samples, we simply replace the Z quantile with a T quantile, as follows: $\hat{\mu} \pm T \cdot Q \cdot \hat{s}$. It comes from a Gosset, or *Student's t-*distribution, which is similar to the normal but has fatter tails.

> A good rule of thumb for when to use T vs Z when both are available, is to **use always T**, since it will approach Z as we gather more data anyway.

Unlike the normal, parametrized by μ, σ, the Gosset is typically centered at zero and with a standard scale, so it has only 1 parameter, the **degrees of freedom**, which, as increase, approximate the function to a standard normal.

The raison d'etre for the Gosset distribution is the following: As seen, this formula approximates a standard normal distribution:

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

But in fact, it would only be a standard normal if our estimated variance S^2 was actually the population variance σ^2. Since it isn't, the result from this normalization is a Gosset distribution. As n increases, the difference doesn't matter, but for low n it can be considerable, and using a standard normal would lead to narrow intervals.

So our t-interval ends up being $\hat{\mu} \pm t_{n-1}\frac{S}{\sqrt{n}}$ where t is the relevant quantile with $n - 1$ degrees of freedom. In the following interactive plot you can see that around 20 it already resembles the normal closely, but for few degrees of freedom it really differs.

```r
library(ggplot2); library(manipulate)
k <- 1000
xvals <- seq(-5, 5, length = k)
myplot <- function(df){
  d <- data.frame(y = c(dnorm(xvals), dt(xvals, df)),
                  x = xvals,
                  dist = factor(rep(c("Normal", "T"), c(k,k))))
  g <- ggplot(d, aes(x = x, y = y))
  g <- g + geom_line(size = 2, aes(colour = dist))
  g
}
manipulate(myplot(mu), mu = slider(1, 30, step = 1))
```

This can be more precisely inspected by plotting the quantiles, as they give the exact difference between the normal and gosset intervals.

Notes about the t-interval:

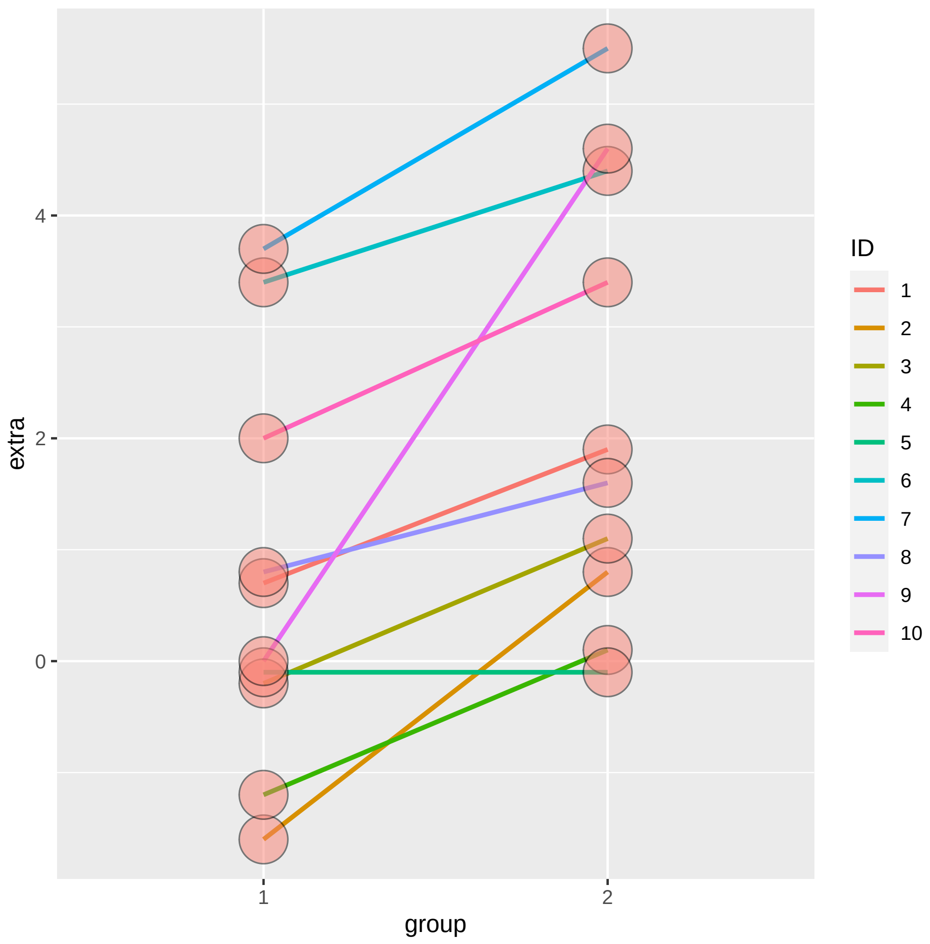
- Although the Gosset assumes that data is iid normal, it is robust to this assumption and "basically whenever the distr. is symmetric and mound shaped, a t-interval will work well".
  - E.g. paired observations: you measure something once, and then same thing days later, and such, are suited for t-intervals.
- For skewed distribution, assumptions aren't favorable anymore:
  - It doesn't make sense to center it at the mean.
  - Taking the log can help.
  - For highly discrete data (e.g. binary, Poisson…), other intervals are available.

Confidence Intervals example

In this example, the sleep duration of 10 people is measured. Then, it is measured again after taking sleep medication. We want to extract a confidence interval for the after-before difference, i.e. how many hours does the sleep change for an individual if they take the medication.

Note that subjects after are same as before, hence we can model it as a **paired t-test**.

The following plot shows the data, the lines link the before-after for the same person:



Since we have few samples and we are computing averages, we will assume that the after-before difference follows a t-distribution. So to find our t-confidence interval:

1. Compute all differences
2. Compute sample mean μ and stddev s of all differences
3. Set $n := 10$
4. The interval is $mu \pm t_{n-1}\frac{s}{\sqrt{n}}$

In R, multiple ways to do it:

```
data(sleep)
head(sleep)
```

```
##   extra group ID
## 1   0.7     1  1
## 2  -1.6     1  2
## 3  -0.2     1  3
## 4  -1.2     1  4
## 5  -0.1     1  5
## 6   3.4     1  6
```

```
g1 <- sleep$extra[1 : 10]; g2 <- sleep$extra[11 : 20]
difference <- g2 - g1
mn <- mean(difference); s <- sd(difference); n <- 10

mn + c(-1, 1) * qt(.975, n-1) * s / sqrt(n)
```

```
## [1] 0.7001142 2.4598858
```

```
t.test(difference)
```

```
##
##  One Sample t-test
##
## data:  difference
## t = 4.0621, df = 9, p-value = 0.002833
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.7001142 2.4598858
## sample estimates:
## mean of x
##      1.58
```

```
# t.test(g2, g1, paired = TRUE)
# t.test(extra ~ I(relevel(group, 2)), paired = TRUE, data = sleep)
```

By our assumptions, the average sleep increase in hours is with 95% confidence within ca. $(0.7, 2.46)$.

Independent group T intervals

Scenario: A/B testing, "performing a randomization in order to balance unobserved covariates that may contaminate your results".

> "**Because you performed this randomization**, it is reasonable to just compare the 2 groups with a t-confidence interval". But we can't used a paired t-test, because there is **no matching of the subjects between the 2 groups**, and they may also have different number of samples.

The confidence interval for **independent groups**(as opposed to paired) X, Y of **same variance** is the following:

$$\bar{Y} - \bar{X} \pm t_{n_x+n_y-2} \cdot S_p \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}$$

The expression is a relatively straightforward adaption of the already known definition for 2 groups, involving one new term:

S_p^2 is the **pooled variance**, defined as follows:

$$S_p^2 = \frac{(n_x - 1)S_x^2 + (n_y - 1)S_y^2}{n_x + n_y - 2}$$

> Assuming same variance between the 2 group populations, pooled variance accounts for $n_x \neq n_y$. The same-variance assumption is reasonable if **we have randomized the groups**. The pooled variance weights more the sampled variance of the group with more samples, so it acts like a weighted interpolation between the 2 groups. In fact, if $n_x = n_y$, the pooled variance is the average of the separate variances.

Example:

- 8 took medication, 21 were control group. Measured blood pressure in mmHg.
- Control group: $\bar{X}_c = 127.44$, $S_c = 18.23$
- Medicated: $\bar{X}_m = 132.86$, $S_m = 15.34$

We want a confidence interval for the difference in averages control and medicated groups.

Since the subjects were randomized, we assume same variance. But we have $n_m! = n_c$. We also have low number of samples. Given this, we apply the formula for t-CI on same-variance independent groups, and a confidence of 95% (i.e. the 0.975 quantile):

$$\begin{aligned} S_p^2 =& \frac{(21-1)18.23^2 + (8-1)15.34^2}{21+8-2} = \frac{8293.867}{27} = 307.1803 \\ & 132.86 - 127.44 \pm t_{21+8-2} \cdot S_p \cdot \sqrt{\frac{1}{21} + \frac{1}{8}} \approx 5.42 \pm 7.282 \cdot t_{27} \\ & = 5.42 \pm 7.282 \cdot 2.051831 = 5.42 \pm 14.9411 = [-9.521, 20.361] \end{aligned}$$

Since **the interval contains zero, it can't be ruled out that zero is the difference in population between the 2 groups**.

The t-student quantile can be retrieved in r as follows:

```
qt(.975, 21)
```

```
## [1] 2.079614
```

Example: Mistakenly treating paired data as group data: We had computed that the 95% t-confidence interval for the difference in averages of the 10 subjects in the sleep dataset was $(0.7, 2.46)$, showing with high confidence that the drug increases sleep time.

> The key here is that the subjects from the second group were the exact same as in the first. If we **mistakenly** treat them as same-variance independent groups and apply this formula, we will get the broader interval $(-0.2, 3.36)$, which includes the zero.

```
# Mistakenly applying the independent-same-variance t-CI
t.test(g2, g1, paired = FALSE, var.equal = TRUE)$conf
```

```
## [1] -0.203874  3.363874
## attr(,"conf.level")
## [1] 0.95
```

```
# Properly treating both groups as paired
t.test(g2, g1, paired = TRUE)$conf
```

```
## [1] 0.7001142 2.4598858
## attr(,"conf.level")
## [1] 0.95
```

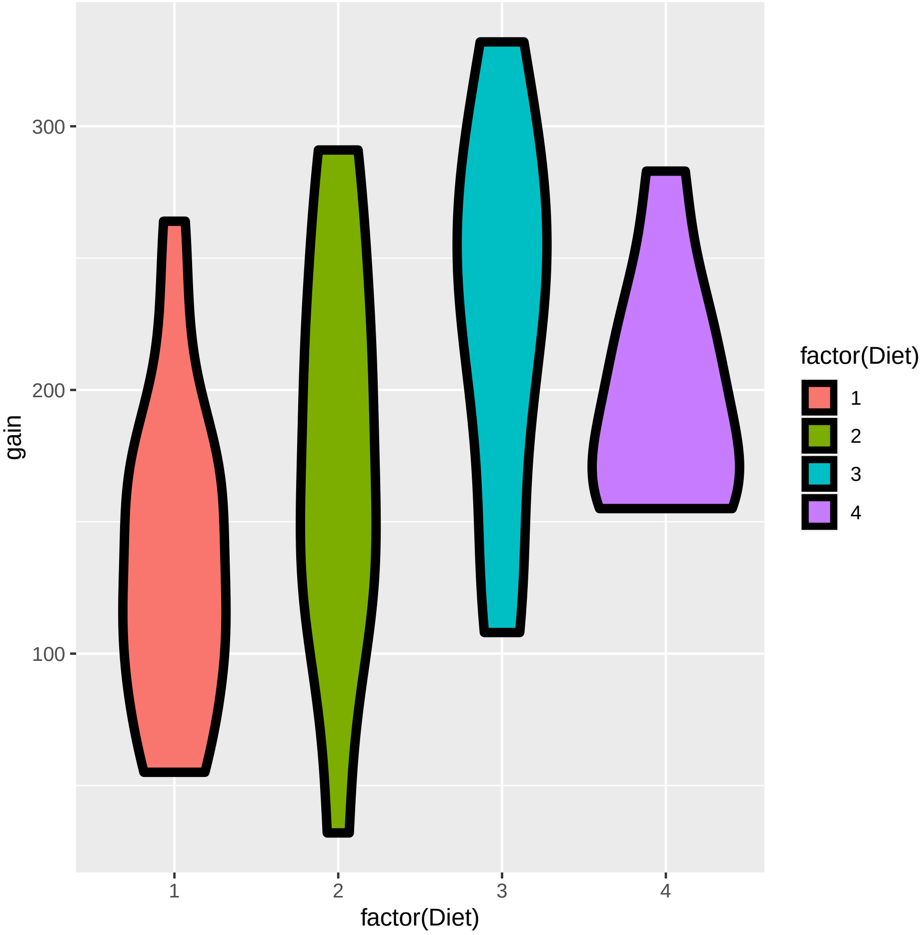
The explanation is that, although the drug does work predominantly well for each subject, there is a lof of variability among subjects (e.g. one sleeps 6 hours, the other 9…). So when *not* paired, what we are comparing both groups as a whole, and all that inter-subject variability lowers the confidence and increases the interval. When taking the per-subject differences, we eliminate all that variability and thus obtain a tighter interval.

Example: Chick weight In this scenario 4 groups of chicks with same initial weight are fed 4 different diets. Each group has different number of chicks, and some seem to gain more weight than others. We want confidence intervals for the comparative between, e.g. groups 1 and 4 (multi-group will be explored on course 2).

The following is a violin plot of the 4gain in weight for the 4 groups (i.e. the thicker a group at a given gain, the more individuals gained that specific weight).

```
##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```



We observe that **our same-variance assumption across groups may be questionable here**. Note how the intervals differ if we assume same or different variance:

```
# Tilde in R: https://stackoverflow.com/a/14976479
wideCW14 <- subset(wideCW, Diet %in% c(1, 4))  # Extract data for diets 1, 4
rbind(
t.test(gain ~ Diet, paired = FALSE, var.equal = TRUE, data = wideCW14)$conf,
t.test(gain ~ Diet, paired = FALSE, var.equal = FALSE, data = wideCW14)$conf
)
```

```
##           [,1]      [,2]
## [1,] -108.1468 -14.81154
## [2,] -104.6590 -18.29932
```

> Exploration about the data is needed to decide if we assume equal variance or not. **When in doubt, assume different variances**: see that the equal interval is tighter than non-equal.

A note on unequal variance

> "If the X and Y observations are iid normal with potentially different μ, σ parameters, the relevant normalized avg. statistic does not follow a t-distribution. **But it can be approximated** by a formula with (potentially) non-integer degrees of freedom, as follows:

$$\bar{Y} - \bar{X} \pm t_{df} \cdot \sqrt{\frac{S_x^2}{n_x} + \frac{S_y^2}{n_y}}$$

$$df := \frac{\left(\frac{S_x^2}{n_x} + \frac{S_y^2}{n_y}\right)^2}{\frac{\left(\frac{S_x^2}{n_x}\right)^2}{n_x - 1} + \frac{\left(\frac{S_y^2}{n_y}\right)^2}{n_y - 1}}$$

So although this scenario doesn't actually follow a Gosset distribution, it can be approximated very closely by this model. In fact,

> "when in doubt, use the unequal variance interval"

Revisiting the blood-pressure example, instead of

$$132.86 - 127.44 \pm 7.282 \cdot t_{27} \approx [-9.521, 20.361]$$

We will have then:

$$\begin{aligned} df :=& \frac{\left(\frac{18.23^2}{21} + \frac{15.34^2}{8}\right)^2}{\frac{\left(\frac{18.23^2}{21}\right)^2}{20} + \frac{\left(\frac{15.34^2}{8}\right)^2}{7}} \approx \frac{(15.825 + 29.414)^2}{12.522 + 123.6} \approx 15.035 \\ tCI =& 132.86 - 127.44 \pm t_{15.035} \cdot \sqrt{\frac{18.23^2}{21} + \frac{15.34^2}{8}} \\ =& 5.42 \pm 2.131 \cdot 6.726 = [-8.91, 19.75] \end{aligned}$$

In R, we don't have to go through the trouble of memorizing the formula and finding the corresponding t-quantile. Given the raw data, simply `t.test(..., var.equal=FALSE)` will do exactly these steps and return the same interval.

Which in R can be computed as follows:

Summarizing:

- t-intervals are some of the handiest in statistics.
- When having single observations, or pairs of observations from which we can extract the difference, t-intervals are highly robust to the underlying assumptions regarding data distributions.
- But there are situations where its preferable to use other procedures (skewed, binary data…) covered in the 2nd course.

Hypothesis testing

> **Hypothesis testing** is concerned with making decisions using data.

1. First, a **null hypothesis** is specified, representing the status quo, and usually labeled H_0 (h-not).
2. H_0 is assumed to be true and statistical evidence is required to reject it in favour of an alternative (research) hypothesis H_a.

The decision bears 2 types of errors:

Truth	Decision	Result
H_0	H_0	Correctly accept null
H_0	H_a	Type I error
H_a	H_a	Correctly reject null
H_a	H_0	Type II error

We can see that decreasing Type I increases Type II, and vice versa. E.g. in court, H_0 is usually that the defendant is innocent. If higher conviction standards are used, more type II errors will happen. If lowered, more type I errors will happen.

Confidence interval example: A population of 100 has an average of 32 events/hour of a disorder, and a standard deviation of 10 events/hour.

So, assuming the normalized mean follows a t-distribution (e.g. by being normally iid distributed) with 99 degrees of freedom, as already seen, the corresponding confidence interval is the following:

$$\mu \pm t_{df-1,(1-\alpha/2)} \cdot \frac{\sigma}{\sqrt{n}} \approx 32 \pm 1.984 \cdot \frac{10}{\sqrt{100}}$$

Which, for 95% confidence, yields approx 32 ± 1.984.

```r
qt(0.975, 99)
```

```
## [1] 1.984217
```

So the interpretation is that, if we were to repeat this procedure for more samples out of the same population, the extracted confidence interval would contain the true mean 95% of the time, under the given assumptions.

Hypothesis testing operates differently (nevertheless there is an equivalence between HT and CI explained later). See next section.

Example of choosing a rejection region

The idea is that we will reject H_0 it if $\bar{X}$ surpasses some threshold C (which takes data variability into account).

> The **constant** C is usually chosen so that the probability of a type I error is some low α (usually 0.05): We *would* like it to be zero, but in that case, we are *never* rejecting it. So we have to settle for a conservative value in rejecting H_0, so that "the type I error probability is *tolerably* low" (usually 5%).

From "Two group testing" video:

> There is an **equivalence between hypothesis testing and confidence intervals**: The 95% set of all possible values for which you fail to reject H_0 (in the double-side test) is a confidence interval for μ. Conversely, if a 95% contains μ_0, then we fail to reject H_0. Therefore, **they can never disagree: the population mean will be outside of our CI exactly as often as our double-side hypothesis will yield a type I error**.

Example So taking the t-CI example from before, now we want to test:

$$H_0 : \quad \mu = 30 H_a : \quad \mu > 30$$

Under H_0, the distribution of the sample mean is $\mathcal{N}(30, 1)$. ** So for our test we want to find C so that

$$P(\bar{X} > C; H_0) = 5\%$$

In other words, since by H_0 we assume the PDF, we are looking for a quantile such that all the density above it integrates to 5%. The reason is that, if we were to repeat our sample experiment many times, 5% of them would end up above. Those would be our type I errors, and we want that to be 5% under our hypothesis. Hence, we guarantee that ratio by rejecting H_0 if our experiment is above C.

For $\mathcal{N}(30, 1)$, the 95th percentile is 31.645. Since our sample data was 32, we reject the hypothesis.

The normalized version of the "greater than" hypothesis testing rule is the following: Reject whenever

$$\frac{\sqrt{n}(\bar{X} - \mu_0)}{s} > Z_{1-\alpha}$$

T tests

Reconsider the example again, but now $n = 16$. So H_0 follows a t-distribution with 15 degrees of freedom. The difference now is that the 95th percentile is not 1.645, but

```r
qt(0.95, 15)
```

```
## [1] 1.75305
```

So the test ends up being

$$\frac{\sqrt{16}(32 - 30)}{10} = 0.8 > 1.7503$$

And as we can see, it fails, meaning that 16 sample points are insufficient to reject H_0.

Two-sided tests If we want to reject H_0 if the mean was too large or too small, i.e.

$$H_a : \mu \neq 30$$

The difference, is that in order to get our 5% guarantee, we need to split it in 2.5% above and same below, so our threshold becomes more demanding. For symmetrical distributions, this is same as if the absolute value of our sample mean is above the 97.5% quantile.

Usually, all this work is done by functions like `t.test`, which output all relevant statistics to perform the test: note that in this case, a t test won't be much different from a z-test, as we have 1078 samples.

```
## Loading required package: MASS
```

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select

## Loading required package: HistData

## Loading required package: Hmisc

## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:dplyr':
##
##     src, summarize

## The following objects are masked from 'package:base':
##
##     format.pval, units

##
## Attaching package: 'UsingR'

## The following object is masked from 'package:survival':
##
##     cancer

##
##  One Sample t-test
##
## data:  father.son$sheight - father.son$fheight
## t = 11.789, df = 1077, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.8310296 1.1629160
## sample estimates:
## mean of x
## 0.9969728
```

the value of `t=11...` is the **t-statistic**, the "estimated difference in the averages between the 2 groups expressed as units in standard error". The greater `abs(t)` the greater evidence against H_0. In the previous example it was 0.8, which was less than our threshold of 1.75, so we didn't reject H_0. Here, the statistic is much bigger, way outside the confidence interval, so we confidently reject that both means are equal.

Two group testing

The rejection rules are the same, but now we want to see if the mean for one group is the same as the mean for another group, i.e.

$$H_0 : \mu_1 = \mu_2$$

Remember the confidence interval formula for same-variance independent groups:

$$\bar{Y} - \bar{X} \pm t_{n_x+n_y-2} \cdot S_p \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}$$

The statistic is then obtained by normalizing as usual:

$$\frac{(\bar{X}_1 - \bar{X}_2) - \mu_0}{\sqrt{\frac{s_1^2}{n_{x_1}} + \frac{s_2^2}{n_{x_2}}}} > Z_{1-\alpha}$$

Example: Exact binomial test Example of hypothesis testing that is not normal or t. A couple has 7 girls out of 8 children, and we want to test if the probability of having a girl is bigger for them:

$$H_0 : p = 0.5 H_a : p > 0.5$$

- What is the relevant rejection region so that the probability of rejecting is less than 5%?

In the previous cases, we took the CDF of the distribution for the corresponding statistic (e.g. average) and found the quantile that satisfied our threshold. Since the binomial is a discrete function, our corresponding statistic (sum of booleans) can only have a discrete set of quantiles.

Since we are performing a "greater than" test, let's observe the survival function of the corresponding binomial:

```
## [1] 1.00000000 0.99609375 0.96484375 0.85546875 0.63671875 0.36328125 0.14453125
## [8] 0.03515625 0.00390625
```

This will tell us, the rate of type I errors that we will obtain, if the threshold is set at that point. I.e. if we reject at zero or more girls, we obviously are going to generate all possible type I errors. And so on.

Note the following:

- In this case there is no way to have a threshold at exactly 5%: the closest is 7 out of 8
- Any alpha below 0.0039 is not possible
- As n increases, we can approximate with a normal distribution
- In this case, we rejected H_0
- Binomial two-sided tests aren't obvious, see further. Note that the exact CIs for the binomial and Poisson are precisely obtained by taking the complementary of the two-sided reject set (called Clopper/Pearson intervals). This is how R gets them.
- P-values help with this, see further

P-Values

From the course slides:

- Most common measure of statistical significance
- Their ubiquity, along with concern over their interpretation and use makes them controversial among statisticians
  - http://warnercnr.colostate.edu/~anderson/thompson1.html
  - Also see *Statistical Evidence: A Likelihood Paradigm* by Richard Royall
  - *Toward Evidence-Based Medical Statistics. 1: The P Value Fallacy* by Steve Goodman
  - The hilariously titled: *The Earth is Round (p < .05)* by Cohen.
- Some positive comments
  - simply statistics
  - normal deviate
  - Error statistics

What is a p-value?

> The probability under H_0 of obtaining evidence (a *test statistic*) as extreme or more as the one obtained through the data.

Suppose H_0 holds: How unusual is the estimate (a.k. "statistic") we got?

1. Define the hypothetical distribution for the statistic under H_0.

2. Measure the statistic from the data (test statistic)
3. Calculate the probability of getting a statistic *as extreme as the measured or more* on the basis of our H_0 distribution.

The p-value is that probability. A low p-value means that either H_0 is false, or we obtained unlikely data, so that the probability of obtaining the measured statistic under the assumption of H_0 is low.

Example: Suppose you get a T-statistic of 2.5 for 15df, testing

$$H_0 : \mu = \mu_0 H_a : \mu > \mu_0$$

What is the p-value for a statistic ≥ 2.5?

```
pt(2.5, 15, lower.tail = FALSE)
```

```
## [1] 0.0122529
```

> The p-value is also known as the **attained significance level**: Imagine we have a p-value of 0.04. Depending on whether our α was 0.05 or 0.01, we would reject or not reject H_0. So if we just provide the p-value, each one can decide whether to reject based on their own threshold.

- Also note: for tests like χ^2 the test is already in a sense two-sided, it doesn't have to be doubled.

P-value examples:

Binom A friend has 8 children, 7 girls:

$$H_0 : p = 0.5 H_a : p > 0.5$$

A reasonable statistic is to count the number of girls. The p-value is the sum of the probability for 7 out of 8 and 8 out of 8, from the distribution $Binom(8, 0.5)$. That is:

```
pbinom(6, 8, 0.5, lower.tail = FALSE)
```

```
## [1] 0.03515625
```

Poisson A hospital has 10 infections per 100 pers*day (i.e. rate of 0.1).

Assume that an infection rate of 0.05 is an important benchmark: above that, some expensive quality procedures must be implemented, but we don't want to implement them just due to noise in data. How can we account for uncertainty?

Assuming a Poisson underlying distribution,

$$\begin{aligned} H_0 &: \lambda = 0.05 \quad \Rightarrow \quad \lambda_0 \cdot 100 = 5 \\ H_a &: \lambda > 0.05 \end{aligned}$$

We want to know the probability of obtaining 10 or more infections for 100 pers*day, assuming H_0:

```r
ppois(9, 5, lower.tail = FALSE)
```

```
## [1] 0.03182806
```

We see that the probability is ca. 3.18%. So depending on our α, we would reject H_0 and implement the procedures.