GEOG 5680
Introduction to R
09: Probabilities and Inference tests

Simon Brewer

Geography Department
University of Utah
Salt Lake City, Utah 84112

simon.brewer@geog.utah.edu

May 04, 2020

Probability

What is this thing called probability?

- Mathematical description of uncertainty
- Tightly linked to statistics for inference
  - Model of population from samples
- Several functions in R for estimating probability
- Also found as part of the inference in test in other functions (e.g. ANOVA, linear models, etc)

Probability

What is this thing called probability?

- Probability shows what *outcomes* might occur given a *model*
  - Given the animal, what are the footprints?

Probability

What is this thing called probability?

- Probability shows what *outcomes* might occur given a *model*
  - Given the animal, what are the footprints?
- Statistics show what *models* might result in a given *outcome*
  - Given the footprints, what is the animal?

Distributions in R

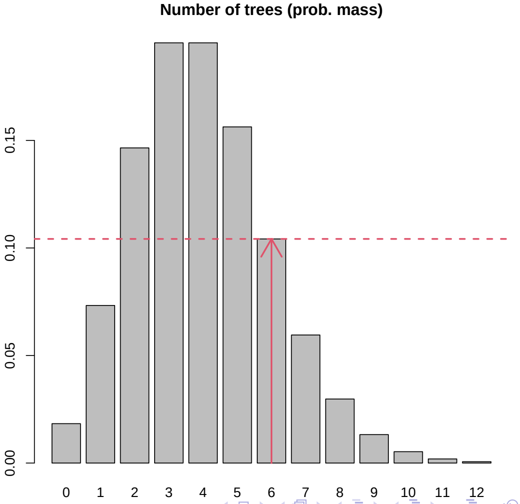
- R comes as standard with approx. 20 well-known probability distribution functions
- Including normal, uniform, binomial, log-normal, beta, gamma, t, F, χ^2 etc
- Add-on packages include approx 100+ extra distributions
- Most distribution come with four functions:
  - `d*` — density functions (e.g. `dnorm()`)
  - `p*` — probability distribution functions (e.g. `pnorm()`)
  - `q*` — quantile functions (e.g. `qnorm()`)
  - `r*` — random number generation (e.g. `rnorm()`)
- Look at examples with Poisson (discrete, count) and normal (continuous)

Poisson distribution

- Count data (λ = mean count)
- `d*`: density function, gives the height of the density curve for a given value
- E.g what is the probability of getting 6 trees in a quadrat?

```
dpois(6,lambda=4)

## [1] 0.1041956
```

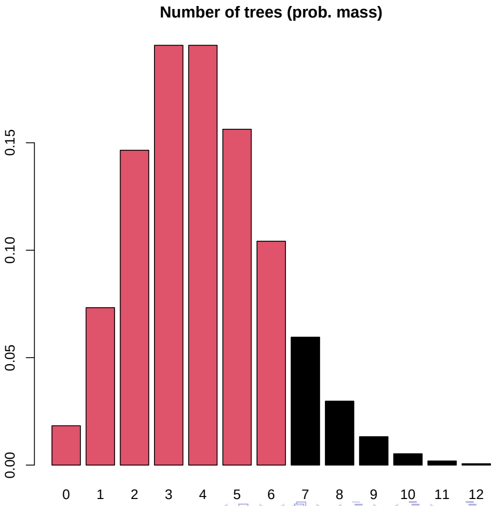


Poisson distribution

- Count data ($\lambda =$ mean count)
- p*: probability dist. function, gives the integral above or below that value
- E.g what is the probability of getting ≤ 6 trees in a quadrat?

```
ppois(6,lambda=4,
      lower.tail = TRUE)

## [1] 0.889326
```

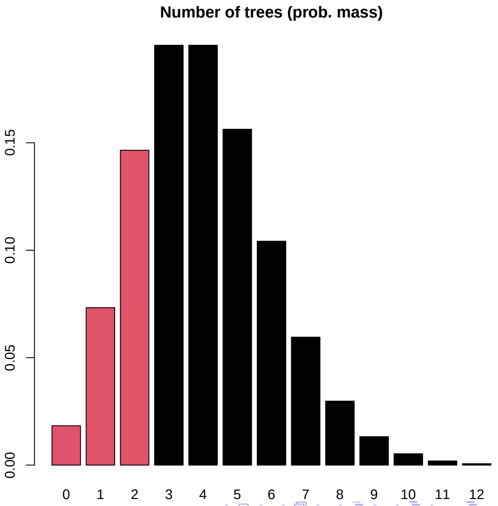


Poisson distribution

- Count data (λ = mean count)
- q*: quantile function, gives the values of X corresponding to a percentile probability
- E.g how many trees do we expect at the 10 percentile of the distribution?

```
qpois(0.1,lambda=4)

## [1] 2
```

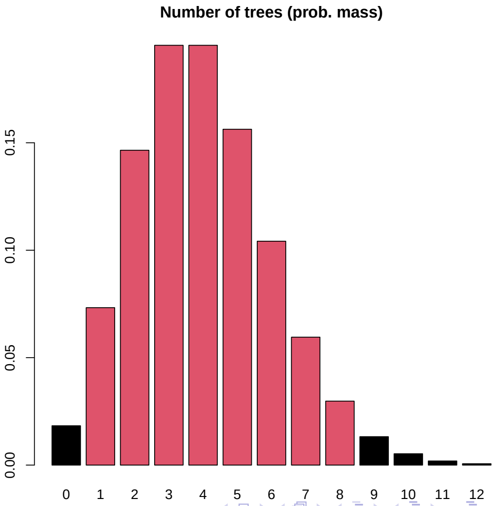


Poisson distribution

- Count data (λ = mean count)
- q*: quantile function, gives the values of X corresponding to a percentile probability
- E.g what is the 95% CI on the number of trees we expect?

```
qpois(c(0.025,0.975),lambda=4)

## [1] 1 8
```



Poisson distribution

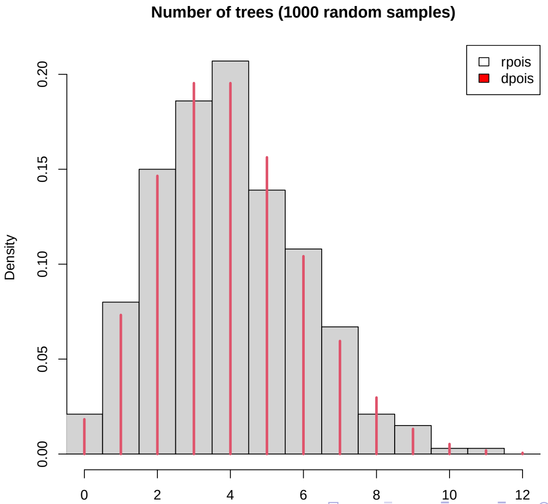
- Count data (λ = mean count)
- `r*`: random function, generates random samples from the distribution
- E.g how many trees might be found in the next four plots?

```
rpois(4,lambda=4)

## [1] 4 4 6 3

rpois(4,lambda=4)

## [1] 4 5 4 2
```

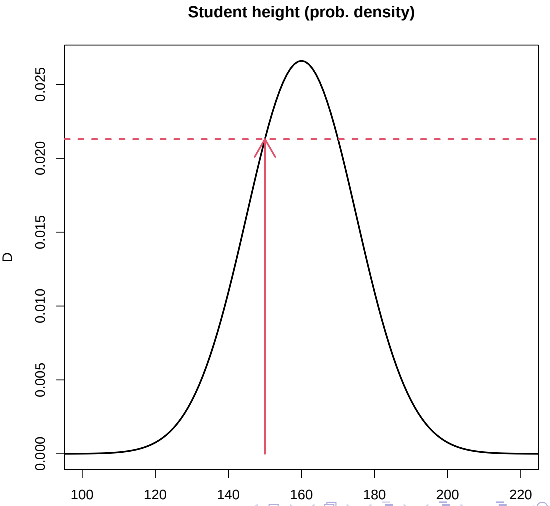


Normal distribution

- Continuous data (μ = mean, σ = std.dev.)
- `d*`: density function, gives the height of the density curve for a given value
- E.g what is the probability density for a height of 150cm?

```r
dnorm(150,mean = 160, sd=15)

## [1] 0.02129653
```

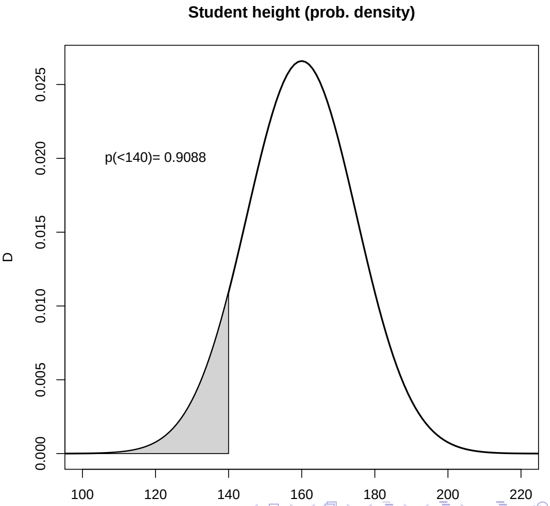


Normal distribution

- Continuous data (μ = mean, σ = std.dev.)
- p*: probability dist. function, gives the integral above or below that value
- E.g what is the probability of a student being smaller than 140cm?

```
pnorm(140, 160, 15,
      lower.tail = TRUE)

## [1] 0.09121122
```

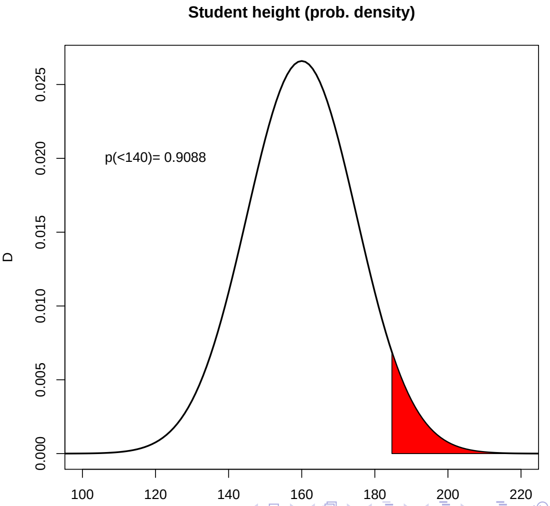


Normal distribution

- Continuous data (μ = mean, σ = std.dev.)
- q*: quantile function, gives the values of X corresponding to a percentile probability
- E.g what cutoff in height gives me the top 5% of students?

```
qnorm(0.95, 160, 15)

## [1] 184.6728
```



Normal distribution

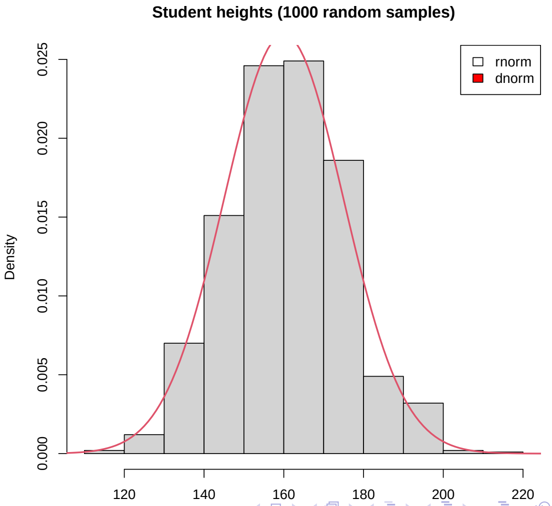
- Continuous data (μ = mean, σ = std.dev.)
- `r*`: random function, generates samples from the distribution
- E.g heights of 3 random students?

```
rnorm(3, 160, 15)

## [1] 145.6133 168.2708 153.2355

rnorm(3, 160, 15)

## [1] 143.2768 159.0669 166.8958
```



Statistical Inference

Statistical Inference and hypothesis testing

- Test some assumptions about a population of interest, using data drawn or sampled from that population
- Compared to descriptive statistics, inference gives significance of a statistical observation
- Examples
  - Do two sets of observations have the same characteristics (mean, variance)?
  - Are two variables correlated among a set of observations?
  - Are observations distributed equally or not?

Student's t-test

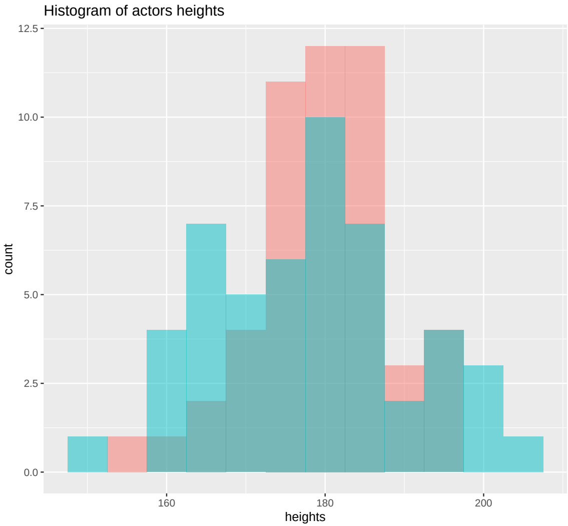
- A t-test is used to compare an observed sample mean (μ_1) to a hypothesized value (μ_0) (one sample t-test)
- Or to compare two sample means (two sample t-test)

$$t = \frac{\mu_1 - \mu_2}{s_{\mu_1 - \mu 2}} \tag{1}$$

- One-tailed ($\mu_1 < \mu_2$ or $\mu_1 > \mu_2$)
- Two-tailed ($\mu_1 \neq \mu_2$)

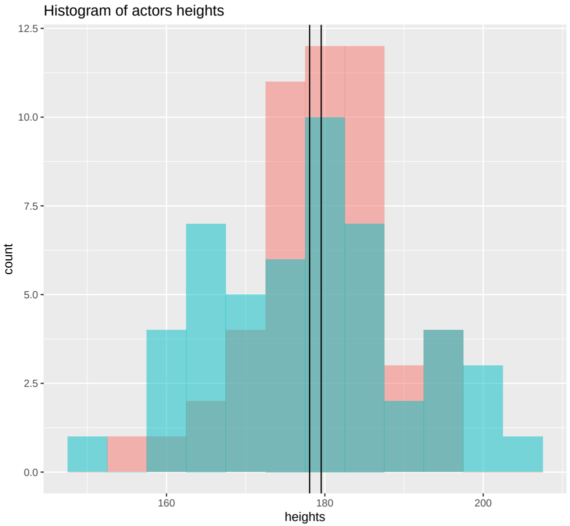
Student's t-test

- Two samples ($n = 50$) of actors who auditioned for the role of Aragorn in Lord of the Rings in two different locations
- Is there a difference in heights?



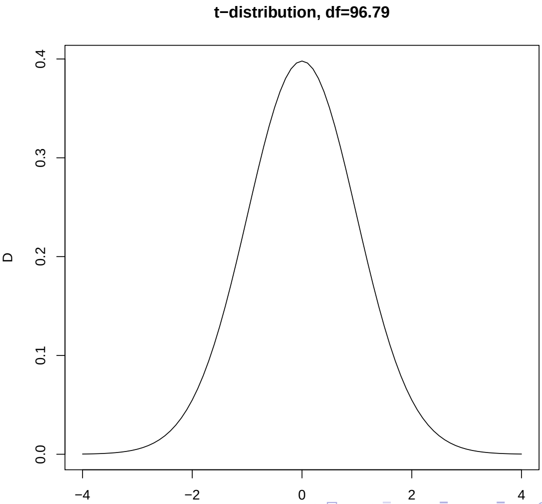
Student's t-test

- Is there a difference in mean height?
- Loc. 1 mean $= 179.54$
- Loc. 2 mean $= 178.07$
- Difference $= 1.46$
- t-statistic $= 0.6886$



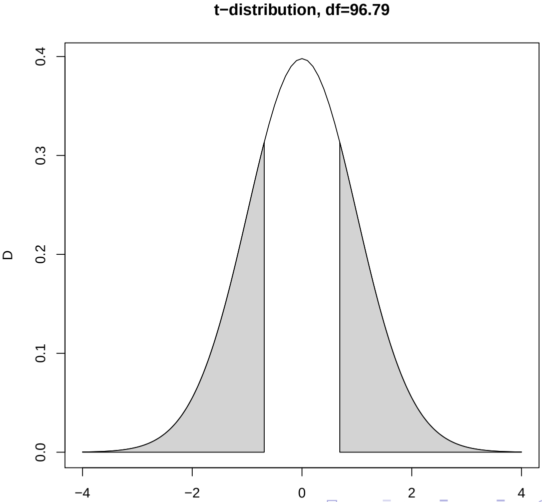
Student's t-test

- Compare t-statistic to t-distribution
- Represents the range of t-statistics expected through normal random variation
- If observed t has a low probability (i.e. in one of the tails), it is less likely to have occured by chance (p-value)



Student's t-test

- Two-tail test:
- The p-value represents the probability that this *difference* (positive or negative) could have occurred by chance
    - p-value is integral of curve $< -|t|$ plus integral of curve $> |t|$
    - p-value $= 0.4929$



Student's t-test

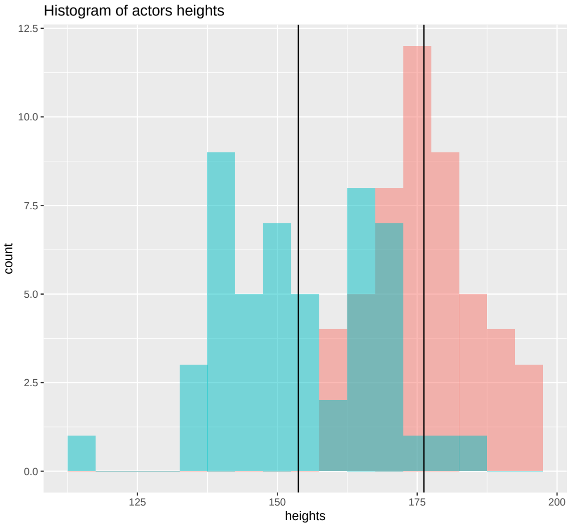
t-test in R using the `t.test()` function:

```
t.test(apop1, apop2, alternative = "two.sided")

##
##  Welch Two Sample t-test
##
## data:  apop1 and apop2
## t = 0.68864, df = 87.394, p-value = 0.4929
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.761618  5.690039
## sample estimates:
## mean of x mean of y
##  179.5392  178.0750
```

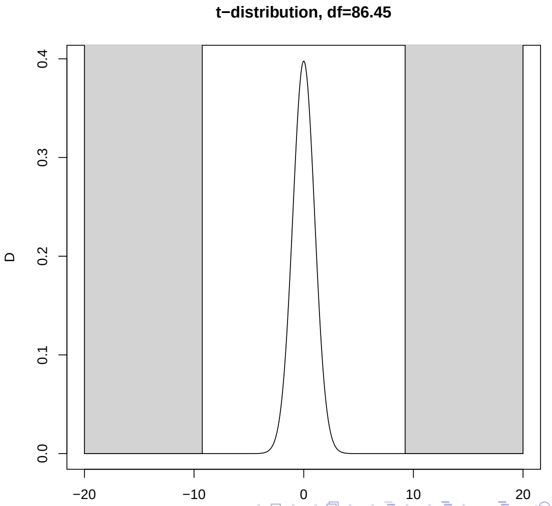
Student's t-test

- If we also had 50 actors who auditioned for Gimli + 50 who auditioned for Aragorn
- Difference = 22.48
- t-statistic = 9.2512



Student's t-test

- The p-value represents the probability that this value (or larger) could have occurred by chance
- Two-tail test:
  - p-value is integral of curve $< -|t|$ plus integral of curve $> |t|$
  - p-value $= 1.4492108 \times 10^{-14}$



Student's t-test

t-test in R using the `t.test()` function:

```
t.test(aragorn, gimli, alternative = "two.sided")

##
##  Welch Two Sample t-test
##
## data:  aragorn and gimli
## t = 9.2512, df = 86.451, p-value = 1.449e-14
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  17.64770 27.30697
## sample estimates:
## mean of x mean of y
##  176.2119  153.7345
```

Student's *t*-test

t-test in R using the `t.test()` function (one-sided):

```
t.test(aragorn, gimli, alternative = "greater")

##
##  Welch Two Sample t-test
##
## data:  aragorn and gimli
## t = 9.2512, df = 86.451, p-value = 7.246e-15
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  18.43762      Inf
## sample estimates:
## mean of x mean of y
##  176.2119  153.7345
```

ANOVA

- What if we have more than two groups?
- If we also have 50 actors who auditioned for Legolas

ANOVA

The F-statistic is used to test for significance in the split of variance:

$$F = \frac{BSS/(t-1)}{ESS/(n-t-1)} \tag{2}$$

- Ratio of how much of the variance is between the groups to how much is within the groups
- Compare to an F-distribution, using degrees of freedom based on the number of groups (t) and the number of observations (n)

ANOVA

- We can use the R function `aov()` to calculate ANOVA for the three groups. Note this uses the model syntax ($\sim$)

```
roles = c(rep("Aragorn",50), rep("Gimli",50), rep("Legolas",50))
heights = c(aragorn, gimli, legolas)
summary(aov(heights ~ roles))

##              Df Sum Sq Mean Sq F value Pr(>F)
## roles         2  42767   21383   181.5 <2e-16 ***
## Residuals   147  17319     118
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Other inference tests

- *F*-test: test if difference in ratio of variance of two samples
  - `var.test()`
- Wilcoxon rank sum test: Non-parametric test for the equality of medians
  - `wilcox.test()`
- Correlation tests: tests of *covariation*
  - `cor.test()`
  - Pearson's vs. Spearman's
- Chi-squared (χ^2) tests: tests of *distribution* and *association*
  - `chisq.test()`