

Handout on Mathematics for EES students - Reminder to statistical testing

Dirk Metzler

October 14, 2022

Contents

1	Basic notations for random variables and distributions	1
2	The binomial distribution	2
3	Normal distribution	6
4	Principle of statistical testing	8
5	Some classical tests	11
5.1	t-test	11
5.2	Analysis of variance (ANOVA)	14
5.3	Chi-square-test	15

1 Basic notations for random variables and distributions

Assume a small population of 100 individuals, and a neutral allele A that has frequency 0.3 in this generation.

What will be the frequency X of A in the next generation?

We don't know, as X is a [random variable](#) .

However, we can ask, for example, for

$\mathbb{E}X = \sum_k k \cdot \Pr(X = k)$, the [expectation value](#) of X , or for

$\Pr(X = 0.32)$, the [probability](#) that X takes a value of 0.32.

Even these values (especially the second one) depend on our [model assumptions](#).

Contents

We start with a simpler Example: Rolling a dice, W is the result of the next trial.

$$\mathcal{S} = \{1, 2, \dots, 6\} \quad \Pr(W = 1) = \dots = \Pr(W = 6) = \frac{1}{6} \quad (\Pr(W = x) = \frac{1}{6} \text{ for all } x \in \{1, \dots, 6\})$$

A **Random Variable** is a result of a random incident or experiment.

The **state space** \mathcal{S} of a random variable is the set of possible values.

The **distribution of a random variable** X assigns to each set $A \subseteq \mathcal{S}$ the **probability** $\Pr(X \in A)$ that X takes a value in A .

In general, we use capitals for random variables (X, Y, Z, \dots) , and small letters (x, y, z, \dots) for (possible) fixed values.

Notations for events

An **event** U like “ X takes a value in A ” is sometimes written with curly brackets:

$$U = \{X \in A\}$$

Stochastic Independence of events

$\Pr(U, V)$: probability that both events U and V take place

$\Pr(U|V)$: conditional probability of U , given that V is known to take place. Note that $\Pr(U|V) = \Pr(U, V) / \Pr(V)$.

Definition 1 (stochastic independence) Two events U and V are *(stochastically) independent* if

$$\Pr(U, V) = \Pr(U) \cdot \Pr(V).$$

Note that $\Pr(U, V) = \Pr(U) \cdot \Pr(V)$ is equivalent to

$$\Pr(U|V) = \Pr(U) \text{ and also to } \Pr(V|U) = \Pr(V)$$

Stochastic Independence of random variables

Definition 2 (stochastic independence) Two random variables X and Y are *(stochastically) independent*, if the identity

$$\Pr(X \in A, Y \in B) = \Pr(X \in A) \cdot \Pr(Y \in B)$$

holds for all (measurable) subsets A and B of the state spaces of X and Y .

Example:

- Tossing two dice: X = result dice 1, Y = result dice 2.

$$\Pr(X = 2, Y = 5) = \frac{1}{36} = \frac{1}{6} \cdot \frac{1}{6} = \Pr(X = 2) \cdot \Pr(Y = 5)$$

2 The binomial distribution

Bernoulli distribution

A **Bernoulli experiment** is an experiment with two possible outcomes “success” and “fail”, or 1 or 0.

Bernoulli random variable X : State space $\mathcal{S} = \{0, 1\}$. Distribution:

$$\Pr(X = 1) = p$$

$$\Pr(X = 0) = 1 - p$$

The parameter $p \in [0, 1]$ is the **success probability**.

Bernoulli distribution

Examples:

- Tossing a coin: 1 and 0 represent “head” and “tail”
- Tossing a drawing pin: 1 and 0 represent “point upward” and “pin down”
- Does the Drosophila have a mutation that causes white eyes? 1 and 0 represent are “yes” and “no”.
- A certain allele on a chromosome: 1 and 0 represent “this allele” and “other allele”

Assume a Bernoulli experiment (for example tossing a coin) with success probability p is repeated n times *independently*.

What is the probability that it...

1. ...always succeeds?

$$p \cdot p \cdot p \cdots p = p^n$$

2. ...always fails?

$$(1 - p) \cdot (1 - p) \cdots (1 - p) = (1 - p)^n$$

3. ...first succeeds k times and then fails $n - k$ times?

$$p^k \cdot (1 - p)^{n-k}$$

4. ...succeeds in total k times and fails the other $n - k$ times?

$$\binom{n}{k} \cdot p^k \cdot (1 - p)^{n-k}$$

Note

$$\binom{n}{k} = \frac{n!}{k! \cdot (n - k)!} = \frac{n \cdot (n - 1) \cdot (n - 2) \cdots (n - k + 1)}{k \cdot (k - 1) \cdot (k - 2) \cdots 3 \cdot 2 \cdot 1}$$

(“ n choose k ”) is the number of possibilities to choose k successes in n trials.

Binomial distribution

Let X be the number of successes in n independent trials with success probability of p each. Then,

$$\Pr(X = k) = \binom{n}{k} p^k \cdot (1 - p)^{n-k}$$

holds for all $k \in \{0, 1, \dots, n\}$ and X is said to be *binomially distributed*, for short:

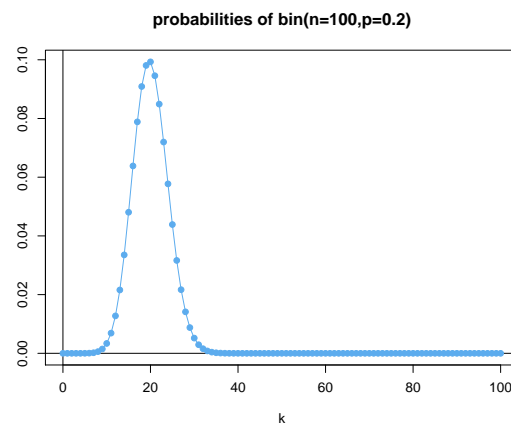
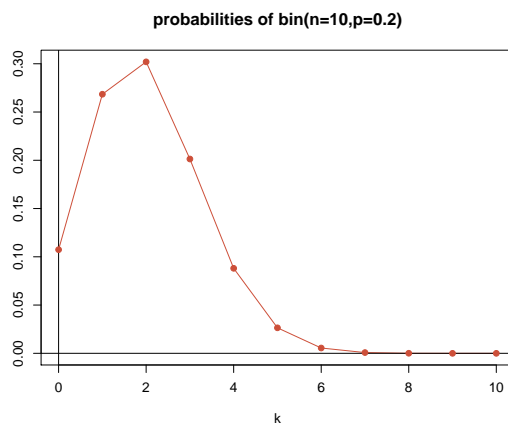
$$X \sim \text{bin}(n, p).$$

Expectation value, variance and standard deviation:

$$\mathbb{E}X = n \cdot p, \quad \text{Var}(X) = n \cdot p \cdot (1 - p), \quad \sigma_X = \sqrt{n \cdot p \cdot (1 - p)}$$

General definition:

$$\text{Var}(X) = \mathbb{E}(X - \mathbb{E}X)^2, \quad \sigma_X = \sqrt{\text{Var}(X)}$$



With the binomial distribution we can treat our initial question

Assume in a small population of $n = 100$ individuals the neutral allele A has a frequency of 0.3.

How probable is it that X , the frequency of A in the next generation is 0.32?

$$\Pr(X = 0.32) = ?$$

We can only answer this on the basis of a probabilistic model, and the answer will depend on how we model the population.

Modeling approach

We make a few simplifying assumptions:

- Discrete generations
- The population is haploid, that is, each individual has exactly one parent in the generation before.
- constant population size $n = 100$

$\Pr(X = 0.32)$ still depends on whether few individuals have many offspring or whether all individuals have similar offspring numbers. $\Pr(X = 0.32)$ is only defined with additional assumptions, e.g.:

- Each individual chooses its parent purely randomly in the generation before.

“purely randomly” means *independent of all others* and *all potential parents with the same probability*.

Our assumptions imply that each individual of the next generations has a probability of 0.3 to obtain allele A , and they get their alleles independently of each other.

Therefore, the number K of individuals who get allele A is binomially distributed with $n = 100$ and $p = 0.3$:

$$K \sim \text{bin}(n = 100, p = 0.3)$$

For $X = K/n$ follows:

$$\begin{aligned} \Pr(X = 0.32) &= \Pr(K = 32) = \binom{n}{32} \cdot p^{32} \cdot (1 - p)^{100-32} \\ &= \binom{100}{32} \cdot 0.3^{32} \cdot 0.7^{68} \approx 0.078 \end{aligned}$$

Genetic Drift

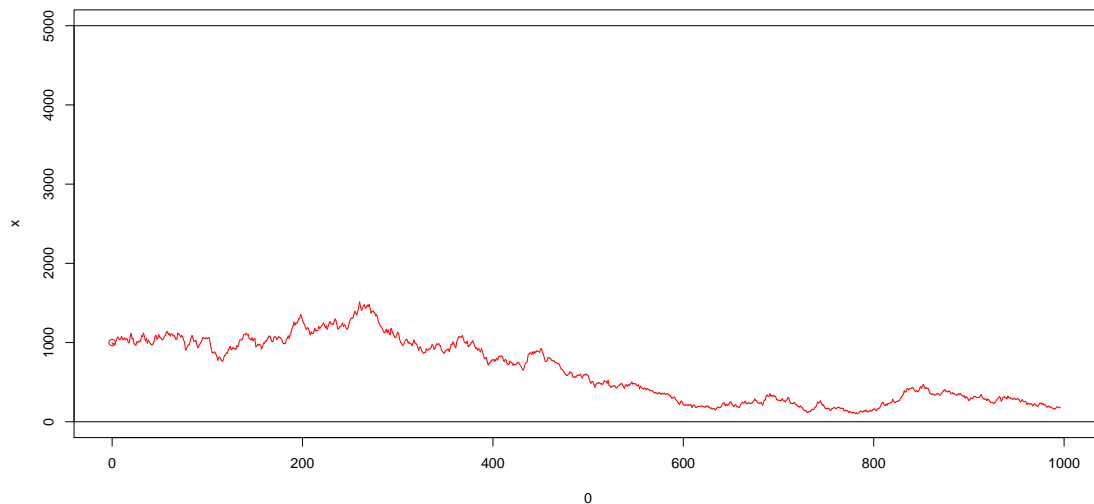
If p is frequency of allele A in the current generation (population size n), the number K of carriers of A in the next generation is $\text{bin}(n, p)$ -distributed and thus satisfies:

$$\mathbb{E}K = n \cdot p \quad \text{and} \quad \sigma_K = \sqrt{n \cdot p \cdot (1 - p)}$$

The allele frequency K/n in the next generation is also a random variable and has the properties:

$$\mathbb{E}(K/n) = n \cdot p/n = p \quad \text{and} \quad \sigma_K = \sqrt{\frac{p \cdot (1 - p)}{n}}$$

Genetic Drift / Wright–Fisher Diffusion



Binomial distribution probabilities in R



The R software is freely available from <https://www.r-project.org/>.
Many like to use R with RStudio:
<https://www.rstudio.com/products/RStudio/>

$$\Pr(K = 32) = \binom{100}{32} \cdot 0.3^{32} \cdot 0.7^{68} \approx 0.078$$

```
> dbinom(32,size=100,p=0.3)
[1] 0.07761057
```

Check by using the formula:

```
> choose(100,32)*0.3^32*(1-0.3)^(100-32)
[1] 0.07761057
```

Binomial distribution probabilities in R

Now assume $B \sim \text{bin}(10, 0.3)$.

$$\Pr(B \leq 2) = \Pr(B = 0) + \Pr(B = 1) + \Pr(B = 2)$$

```
> pbinom(2,size=10,p=0.3)
[1] 0.3827828
```

Again, check this by step-wise calculation:

```
> dbinom(0:2,size=10,p=0.3)
[1] 0.02824752 0.12106082 0.23347444
> sum(dbinom(0:2,size=10,p=0.3))
[1] 0.3827828
```

Binomial distribution probabilities in R

Still assume $B \sim \text{bin}(10, 0.3)$.

$$\Pr(B > 8) = \Pr(B = 9) + \Pr(B = 10)$$

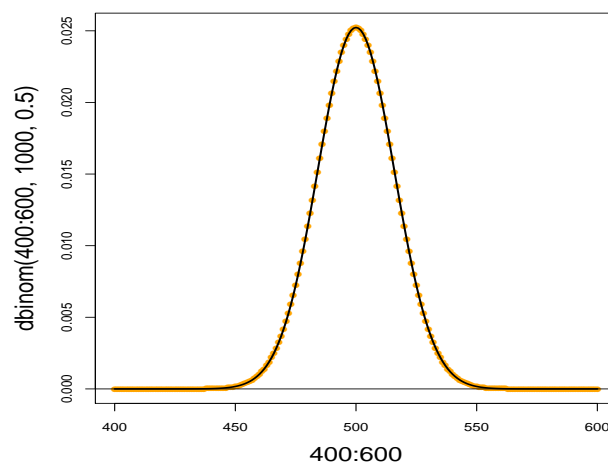
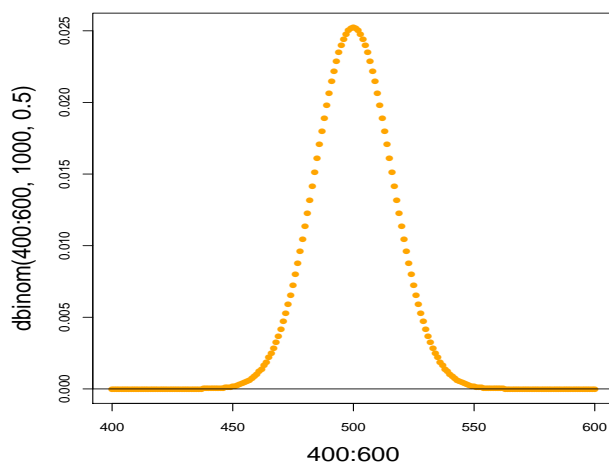
```
> pbinom(8,size=10,p=0.3,lower.tail=FALSE)
[1] 0.0001436859
```

Again, check this by step-wise calculation:

```
> dbinom(9:10,size=10,p=0.3)
[1] 1.37781e-04 5.90490e-06
> sum(dbinom(9:10,size=10,p=0.3))
[1] 0.0001436859
```

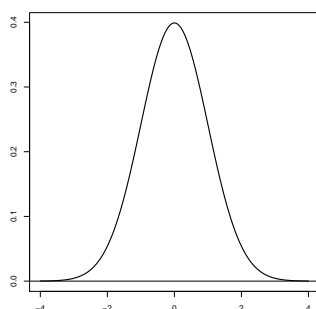
3 Normal distribution

A binomial distribution with large n looks like a normal distribution:



Density of the standard normal distribution

A random variable Z with the density $f(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{x^2}{2}}$ is called *standard-normally distributed*.



“Gaussian bell-curve”

for short:
 $Z \sim \mathcal{N}(0, 1)$

$$\mathbb{E}Z = 0$$

$$\text{Var } Z = 1$$

If Z is $\mathcal{N}(0, 1)$ distributed, then $X = \sigma \cdot Z + \mu$ is normally distributed with mean μ and variance σ^2 , for short:

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

X has the density

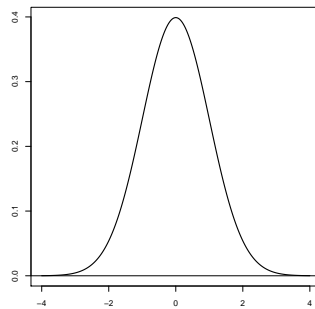
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Question: How to compute $\Pr(Z = 5)$?

Answer: For each $x \in \mathbb{R}$ we have $\Pr(Z = x) = 0$ (Area of width 0)

example: density of the standard normal distribution:

```
> plot(dnorm,from=-4,to=4)
```

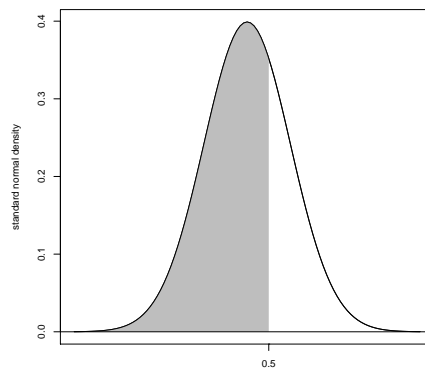


```
> dnorm(0) [1] 0.3989423 > dnorm(0,mean=1,sd=2) [1] 0.1760327
```

example: Computing probabilities: Let $Z \sim \mathcal{N}(\mu = 0, \sigma^2 = 1)$ be standard normally distributed

$\Pr(Z < a)$ can be computed in R by `pnorm(a)`

```
> pnorm(0.5) [1] 0.6914625
```



example: Computing probabilities: Let $Z \sim \mathcal{N}(\mu = 5, \sigma^2 = 2.25)$.

Computing $\Pr(Z \in [3, 4])$:

$$\Pr(Z \in [3, 4]) = \Pr(Z < 4) - \Pr(Z < 3)$$

```
> pnorm(4,mean=5,sd=1.5)-pnorm(3,mean=5,sd=1.5) [1] 0.1612813
```

Normal approximation

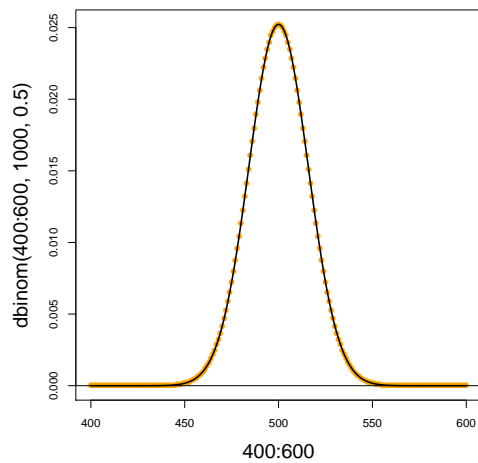
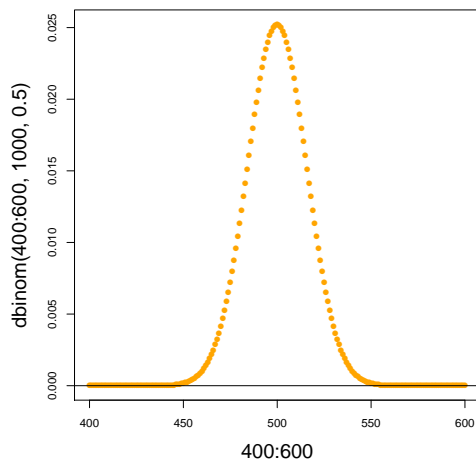
For large n and p which are not too close to 0 or 1, we can approximate the binomial distribution by a normal distribution with the corresponding mean and variance.

If $X \sim \text{bin}(n, p)$ and $Z \sim \mathcal{N}(\mu = n \cdot p, \sigma^2 = n \cdot p \cdot (1 - p))$, we get

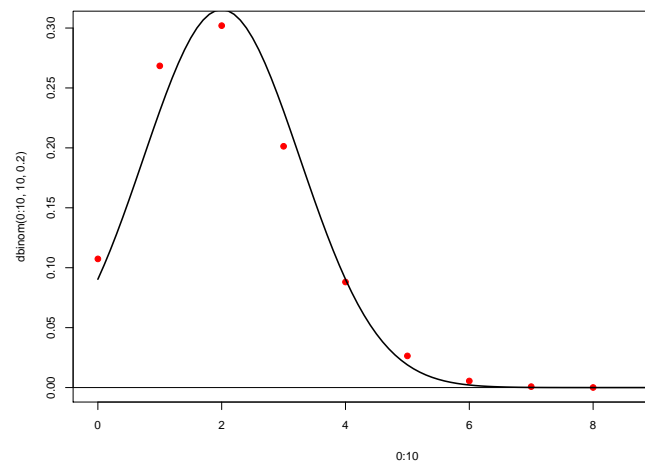
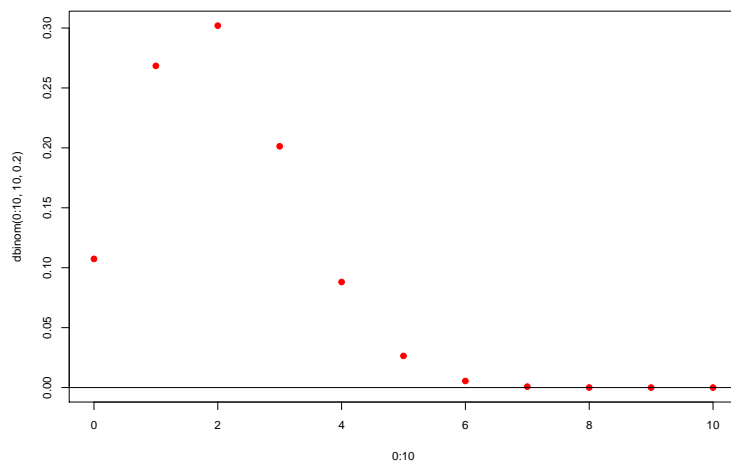
$$\Pr(X \in [a, b]) \approx \Pr(Z \in [a, b])$$

(rule of thumb: Usually okay if $n \cdot p \cdot (1 - p) \geq 9$)

$n = 1000$, $p = 0.5$, $n \cdot p \cdot (1 - p) = 250$



$$n = 10, p = 0.2, n \cdot p \cdot (1 - p) = 1.6$$



4 Principle of statistical testing

Cats or dogs?

- We asked a representative sample of 320 LMU students whether they like cats better than dogs or vice versa.
- 168 said they prefer dogs, 132 preferred cats and 20 were undecided.
- Can we conclude that there is significant evidence that a majority of LMU students prefer dogs over cats?

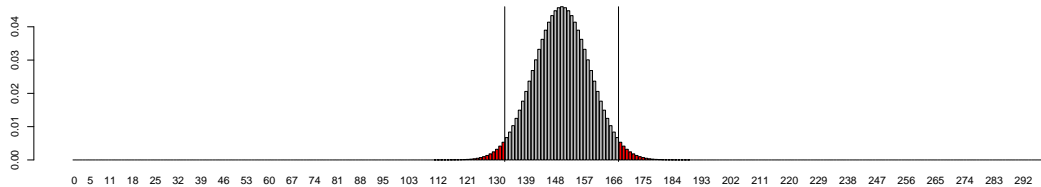
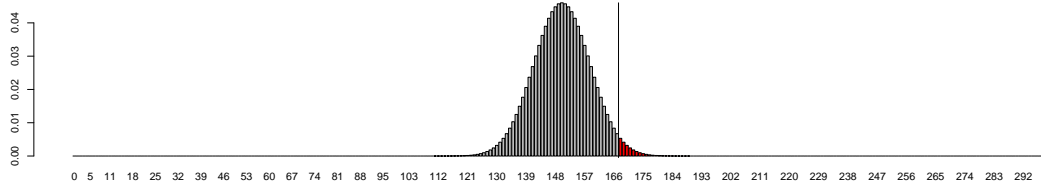
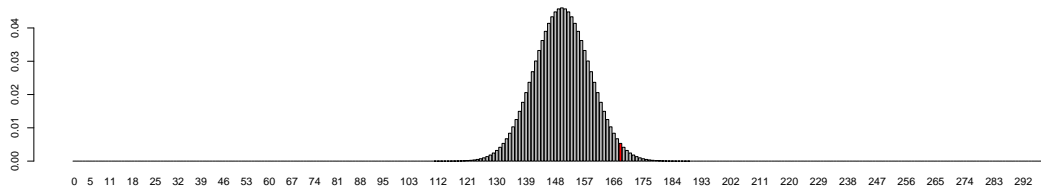
(Of course the following data are purely hypothetical and this survey was never made – at least as far as I know.)

Among the $n = 300$ students who had a preference at all, $K = 168$ preferred dogs over cats. Significantly different from 150?

Null hypothesis: Among the LMU students with a preference, exactly half like dogs better than cats.

If the null hypothesis is true, the number K in a study as above is $\text{bin}(n = 300, p = 0.5)$ -distributed.

How improbable is a deviation of 18 from the $n \cdot p = 150$ if the null hypothesis is true?



<code>dbinom(168,300,p=0.5)</code>	<code>pbinom(167,300,p=0.5,lower.tail=FALSE)</code>
0.005318873	0.02156425
<code>dbinom(150,300,p=0.5)</code>	<code>pbinom(132,300,p=0.5) + pbinom(167,300,p=0.5,lower.tail=FALSE)</code>
0.04602751	0.0431285

```
pbinom(167,300,p=0.5,lower.tail=FALSE)
0.02156425
pnorm(167,mean=150,sd=sqrt(75),lower.tail=FALSE)
0.02482361
pnorm(168,mean=150,sd=sqrt(75),lower.tail=FALSE)
0.01883346
sum(dnorm(168:300,mean=150,sd=sqrt(75)))
0.02159596
```

Statistical testing

- We want to argue that some deviation in the data is not just random.
- To this end we first specify a **null hypothesis** H_0 , i.e. we define, what “just random” means.
- Then we try to show: If H_0 is true, then a deviation that is *at least* as large as the observed one, is very improbable.
- If we can do this, we reject H_0 .
- How we measure **deviation**, must be clear *before* we see the data.

Statistical Testing: Important terms

null hypothesis H_0 : says that what we want to substantiate is not true and anything that looks like evidence in the data is just random. We try to reject H_0 .

significance level α : If H_0 is true, the probability to falsely reject it, must be $\leq \alpha$ (often $\alpha = 0.05$).

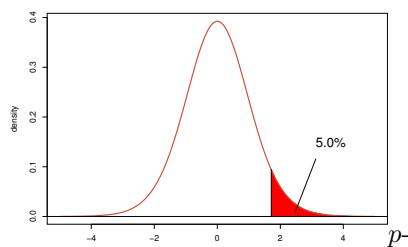
test statistic : measures how far the data deviates from what H_0 predicts into the direction of our alternative hypothesis.

p value : Probability that, if H_0 is true, a dataset leads to a test statistic value that is as least as extreme as the observed one.

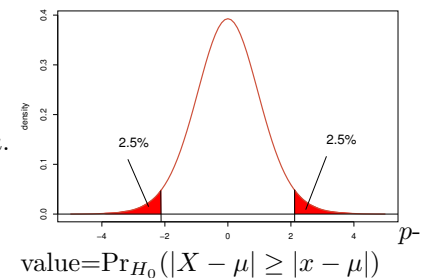
- We reject the null hypothesis H_0 if the p value is smaller than α .
- Thus, if H_0 is true, the probability to (falsely) reject it is α (not the p value).
- This entails that a researcher who performs many tests with $\alpha = 0.05$ on complete random data (i.e. where H_0 is always true), will falsely reject H_0 in 5% of the tests.
- Therefore it is a severe violation of academic soundness to perform tests until one shows significance, and to publish only the latter.

Testing two-sided or one-sided?

We observe a value of x that is much larger than the H_0 expectation value μ .



value = $\Pr_{H_0}(X \geq x)$



Important

The decision between one-sided and two-sided must not depend on the concrete data that are used in the test. More generally: If \mathcal{A} is the event that will lead to the rejection of H_0 , (if it occurs) then \mathcal{A} must be defined without being influenced by the data that is used for testing.

If H_0 is rejected on the 5%-level, which of the following statements is true?

- The null hypothesis is wrong. ~~The null hypothesis is wrong.~~
- H_0 is wrong with a probability of 95%. ~~H_0 is wrong with a probability of 95%.~~
- If H_0 is true, you will see such an extreme event only in 5% of the data sets. If H_0 is true, you will see such an extreme event only in 5% of the data sets. ✓

If the test did not reject H_0 , which of the following statements are true?

- We have to reject the alternative H_1 . ~~We have to reject the alternative H_1 .~~

- H_0 is true. ~~H_0 is true~~
- H_0 is probably true. ~~H_0 is probably true.~~
- It is safe to assume that H_0 was true. ~~It is safe to assume that H_0 was true.~~
- H_0 is compatible with the data, at least with respect to the test statistic. H_0 is compatible with the data, at least with respect to the test statistic. ✓

5 Some classical tests

5.1 t-test

one-sample t-test

Data: values X_1, X_2, \dots, X_n with mean \bar{X} and variance $s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$.

Required: data sampled independently from (approximately) a normal distribution with unknown mean μ and unknown variance σ^2 .

H_0 : $\mu = \mu_0$

Test statistic:

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

Note that s/\sqrt{n} is the **standard error (of the mean; SEM)**.

Distribution of t under H_0 : Student's t distribution with $(n - 1)$ degrees of freedom (df).

paired two-sample t-test

Data: pairs of values $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$.

Required: data sampled independently from (approximate) normal distributions with unknown means μ_X and μ_Y .

H_0 : $\mu_X = \mu_Y$

Test: one-sample t-test with data $X_1 - Y_1, X_2 - Y_2, \dots, X_n - Y_n$ with null hypothesis $\mu = 0$.

two-sample t-test assuming equal variances

Data: samples X_1, X_2, \dots, X_n and Y_1, Y_2, \dots, Y_m with means \bar{X} and \bar{Y} and pooled sample variance

$$s^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2 + \sum_{j=1}^m (Y_j - \bar{Y})^2}{m + n - 2}$$

Required: data sampled independently from (approximate) normal distributions with (unknown) means μ_X and μ_Y and (unknown) equal variance σ^2 .

H_0 : $\mu_X = \mu_Y$

Test statistic:

$$t = \frac{\bar{X} - \bar{Y}}{s \cdot \sqrt{\frac{1}{n} + \frac{1}{m}}}$$

Distribution of t under H_0 : Student's t statistic with df = $n + m - 2$

Welch's t-test

Data: samples X_1, X_2, \dots, X_n and Y_1, Y_2, \dots, Y_m with means \bar{X} and \bar{Y} and sample variances

$$s_X^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1} \quad s_Y^2 = \frac{\sum_{j=1}^m (Y_j - \bar{Y})^2}{m-1}$$

Required: data sampled independently from (approximate) normal distributions with (unknown) means μ_X and μ_Y and (unknown) variances σ_X^2 and σ_Y^2 .

H_0 : $\mu_X = \mu_Y$

Test statistic:

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}}$$

Distribution of t under H_0 : Approximately (!) Student's t statistic with complicated formula for df.

```
> x <- c(2.327429, 2.723787, 4.870450, 3.084610, 3.155145, 5.058078, 3.553099,
        1.481927, 2.175777, 2.465206)
```

```
> x
[1] 2.327429 2.723787 4.870450 3.084610 3.155145 5.058078 3.553099 1.481927
[9] 2.175777 2.465206
```

```
> t.test(x,mu=5)
```

One Sample t-test

```
data: x
t = -5.2784, df = 9, p-value = 0.0005082
alternative hypothesis: true mean is not equal to 5
95 percent confidence interval:
 2.270796 3.908306
sample estimates:
mean of x
 3.089551
```

```
> y
[1] 3.072246 3.295750 5.450604 3.606747 3.543977 5.915461 4.152670 1.588603
[9] 2.816048 2.870647
> t.test(x-y)
```

One Sample t-test

```
data: x - y
t = -8.2513, df = 9, p-value = 1.727e-05
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
-0.6902430 -0.3932058
sample estimates:
mean of x
-0.5417244
```

```
> t.test(x,y,paired=TRUE)
```

Paired t-test

data: x and y

```

t = -8.2513, df = 9, p-value = 1.727e-05
alternative hypothesis: true difference in means is
not equal to 0
95 percent confidence interval:
 -0.6902430 -0.3932058
sample estimates:
mean of the differences
      -0.5417244

> t.test(x,y,var.equal=TRUE)

```

Two Sample t-test

```

data: x and y
t = -0.9995, df = 18, p-value = 0.3308
alternative hypothesis: true difference in means is
not equal to 0
95 percent confidence interval:
 -1.6804089  0.5969601
sample estimates:
mean of x mean of y
  3.089551  3.631275

> t.test(x,y)

```

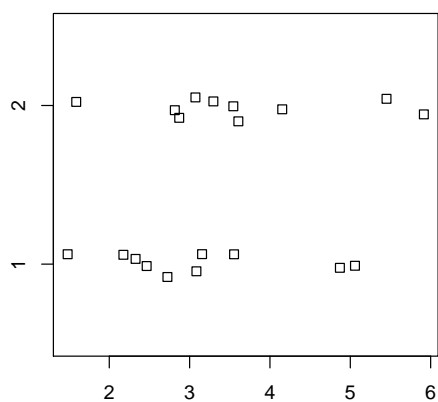
Welch Two Sample t-test

```

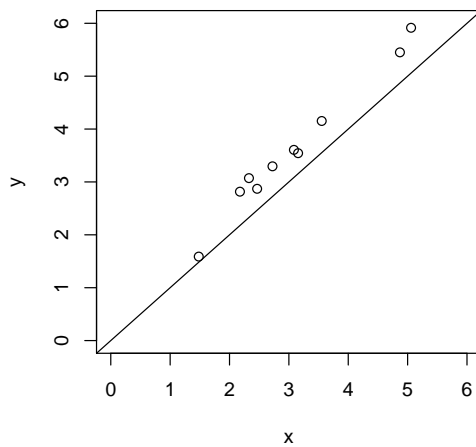
data: x and y
t = -0.9995, df = 17.792, p-value = 0.331
alternative hypothesis: true difference in means is
not equal to 0
95 percent confidence interval:
 -1.681364  0.597915
sample estimates:
mean of x mean of y
  3.089551  3.631275

```

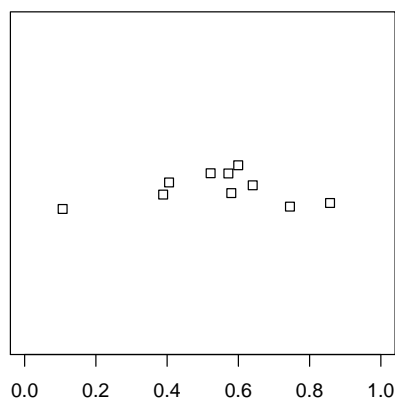
```
stripchart(list(x,y),ylim=c(0.5,2.5),method="jitter")
```



```
plot(x,y,xlim=c(0,6),ylim=c(0,6))
abline(a=0,b=1)
```



```
stripchart(y-x,method="jitter",xlim=c(0,1))
```



5.2 Analysis of variance (ANOVA)

(one-way) anova

Like unpaired t-test with equal variances, but with more than two groups.

Data: For each group g of G groups a number J_g of values $X_{g,1}, \dots, X_{g,J_g}$. Let \overline{X}_g be the mean in group g and \overline{X} be the mean of all values. Let $n = \sum_{g=1}^G J_g$ be the total number of values.

Required: Data sampled independently; within all groups normally distributed with the same variance.

H_0 : The normal distributions of the groups have the same mean.

Test statistic:
$$F = \frac{\sum_{g=1}^G J_g (\overline{X}_g - \overline{X})^2 / (G-1)}{\sum_{g=1}^G \sum_{j=1}^{J_g} (X_{g,j} - \overline{X}_g)^2 / (n-G)}$$

Distribution: If H_0 holds, F is Fisher distributed with $G - 1$ and $n - G$ degrees of freedom.

```

> d <- data.frame(treat,obs)
> d
  treat  obs
1     A 0.69
2     A 0.55
3     A -0.06
4     B 3.69
5     B 3.62
6     B 3.31
7     C 0.79
8     C -0.27
9     C 3.17
10    C 2.21

> mod <- lm(obs~treat,data=d)
> anova(mod)
Analysis of Variance Table

Response: obs
      Df Sum Sq Mean Sq F value Pr(>F)
treat   2 15.4324   7.7162   7.3715 0.01893 *
Residuals  7  7.3274   1.0468
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

```

5.3 Chi-square-test

```

> M <- matrix(c(10,13,12, 4,
+              7,24, 8,11,
+              12,43,36,42
+              ),byrow=TRUE,nrow=3,
+              dimnames=list(c("EES","Neuro","MolBiol"),
+                             c("Steak","Pasta","Pizza","Burger")))
> M
      Steak Pasta Pizza Burger
EES      10    13    12     4
Neuro     7    24     8    11
MolBiol   12    43    36    42

```

Null hypothesis: what a student chose for lunch yesterday was independent of his or her study program.

```
> chisq.test(M)
```

Pearson's Chi-squared test

```
data:  M
X-squared = 17.011, df = 6, p-value = 0.009241
```

Chi-square-test of independence/homogeneity

Data: Contingency table with n rows and m columns; let O_{ij} the (integer) number in row i and column j , $R_i = \sum_j O_{ij}$, $C_j = \sum_i O_{ij}$, $S = \sum_i \sum_j O_{ij}$.

H_0 : Rows are independent of columns, that is, same distribution in all rows (or columns). Conditioned on all R_i and C_j , the expectation of O_{ij} is $E_{ij} = R_i \cdot C_j / S$.

Test statistic: $X^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$

Distribution of X^2 under H_0 : approx. χ^2 -distributed with $(n - 1) \cdot (m - 1)$ degrees of freedom.