Advanced Quantitative Methods: Hypothesis testing

Johan A. Elkink

University College Dublin

12 February 2013

Outline

Bernoulli trial

"An experiment in which s trials are made of an event, with probability p of success in any given trial."

(Weisstein, Eric W. "Bernoulli Trial." http://mathworld.wolfram.com/BernoulliTrial.html)

Binomial distribution

"The (...) probability distribution (...) of obtaining exactly n successes out of N Bernoulli trials."

Binomial distribution

"The (...) probability distribution (...) of obtaining exactly n successes out of N Bernoulli trials."

$$P(n|N) = \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n}$$

(Weisstein, Eric W. "Binomial Distribution." http://mathworld.wolfram.com/BinomialDistribution.html)

Binomial distribution

$$P(n|N) = \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n}$$

$$\lim_{N\to\infty} p(n) = \frac{1}{\sqrt{2\pi Npq}} e^{-\frac{(n-Np)^2}{2Npq}}$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(n-\bar{n})^2}{2\sigma^2}}$$

Binomial distribution

$$P(n|N) = \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n}$$

$$\lim_{N \to \infty} p(n) = \frac{1}{\sqrt{2\pi Npq}} e^{-\frac{(n-Np)^2}{2Npq}}$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(n-\bar{n})^2}{2\sigma^2}}$$

i.e. the **limiting distribution** of the binomial distribution is the **normal distribution**, with $\sigma^2 \equiv Npq$.

(Weisstein, Eric W. "Binomial distribution." http://mathworld.wolfram.com/binomialdistribution.html)

Normal distribution

Also called **Gaussian distribution**

Normal distribution

Also called **Gaussian distribution**, but Gauss did not invent it.

(Davidson & MacKinnon 1999: 130-135)

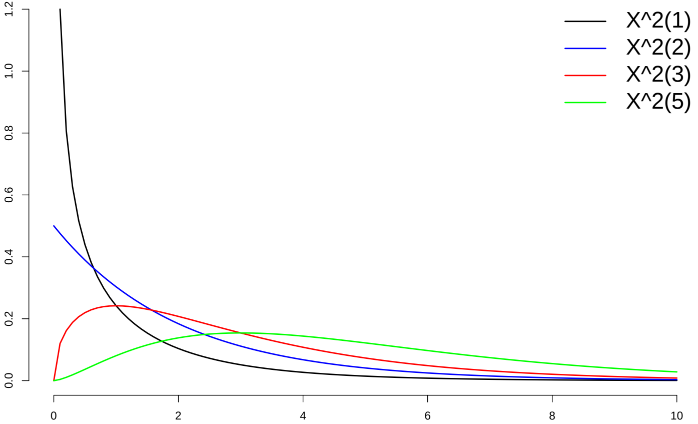
χ^2-distribution

The sum of squares of r independent standard normal distributions, is distributed chi-squared with r degrees of freedom, i.e. if $x \sim N(0,1)$, then:

$$\sum_i^r x_i^2 \sim \chi^2(r)$$

(Weisstein, Eric W. "Chi-squared distribution." http://mathworld.wolfram.com/chi-squareddistribution.html)

χ^2-distribution



F-distribution

If

$$x \sim \chi^2(m)$$
$$y \sim \chi^2(n)$$

with x, y independent

F-distribution

If

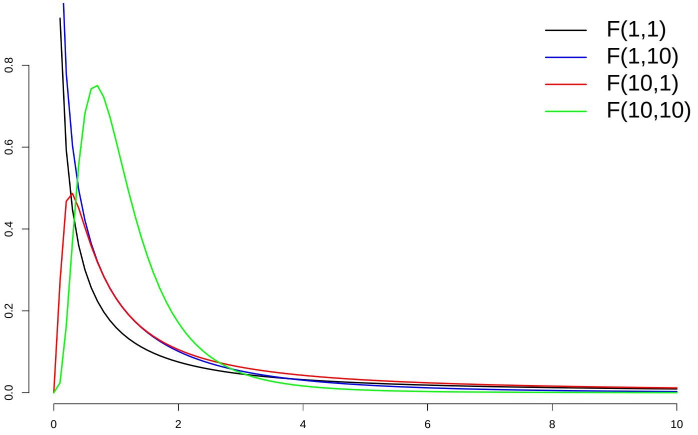
$$x \sim \chi^2(m)$$
$$y \sim \chi^2(n)$$

with x, y independent, then

$$\frac{x/m}{y/n} \sim F(m,n)$$

has an F-distribution with m and n degrees of freedom.

F-distribution



t-distribution

If

$$x \sim N(0,1)$$
$$y \sim \chi^2(r)$$

with x, y independent

t-distribution

If

$$x \sim N(0,1)$$
$$y \sim \chi^2(r)$$

with x, y independent, then

$$\frac{x}{\sqrt{y/r}} \sim t(r)$$

has a t-distribution with r degrees of freedom.

t-distribution

Imagine we have a sample of size n and we calculate the sample mean of x:

$$\bar{x} = \frac{1}{n}\sum_{i}^{n} x_i$$

with variance estimator

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i}^{n}(x_i - \bar{x})^2$$

t-distribution

Imagine we have a sample of size n and we calculate the sample mean of x:

$$\bar{x} = \frac{1}{n}\sum_{i}^{n} x_i$$

with variance estimator

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i}^{n}(x_i - \bar{x})^2$$

then $\hat{\sigma}^2 \sim \chi^2(n-1)$ (because $\hat{\sigma}^2$ is a sum of squares and the use of deviations from the mean removes one degree of freedom).

t-distribution

$\bar{x}$ is a sum of normally distributed values, so is itself normally distributed; $\hat{\sigma}^2$ has a χ^2 distribution, so

$$t(n) = \frac{\bar{x} - \mu}{\sqrt{\hat{\sigma}^2 / n}}$$

has a t-distribution with $n - 1$ degrees of freedom.

t-distribution

$\bar{x}$ is a sum of normally distributed values, so is itself normally distributed; $\hat{\sigma}^2$ has a χ^2 distribution, so

$$t(n) = \frac{\bar{x} - \mu}{\sqrt{\hat{\sigma}^2/n}}$$

has a t-distribution with $n-1$ degrees of freedom.

As n increases, the t-distribution approaches a normal distribution. The t-distribution is the approximation for the normal distribution when n is small and σ^2 unknown.

t-distribution

$\bar{x}$ is a sum of normally distributed values, so is itself normally distributed; $\hat{\sigma}^2$ has a χ^2 distribution, so

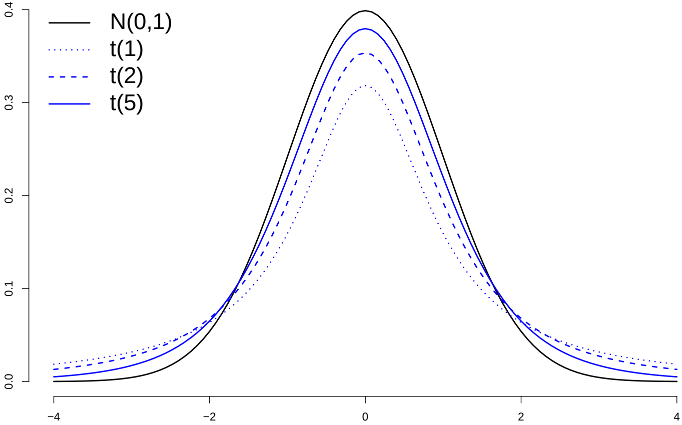
$$t(n) = \frac{\bar{x} - \mu}{\sqrt{\hat{\sigma}^2 / n}}$$

has a t-distribution with $n - 1$ degrees of freedom.

As n increases, the t-distribution approaches a normal distribution. The t-distribution is the approximation for the normal distribution when n is small and σ^2 unknown.

(The $t(1)$ distribution is also called the **Cauchy distribution**.)

t-distribution



Outline

i.i.d.

We make three assumptions about our data to proceed:

- The observations are **independent**
- The observations are **identically distributed**
- The population has a finite mean and a finite variance

A variable for which the first two assumptions hold is called **iid**.

Independent observations

Intuitively: the value for one case does not affect the value for another case on the same variable.
More formally: $P(x_1 \cap x_2) = P(x_1)P(x_2)$.

Independent observations

Intuitively: the value for one case does not affect the value for another case on the same variable.
More formally: $P(x_1 \cap x_2) = P(x_1)P(x_2)$.

Examples of dependent observations:

- grades of students in different classes;
- stock values over time;
- economic growth in neighbouring countries.

Identically distributed

All the observations are drawn from the same **random variable** with the same **probability distribution**.

Identically distributed

All the observations are drawn from the same **random variable** with the same **probability distribution**.

An example where this is not the case would generally be panel data. E.g. larger firms will have larger variations in profits, thus their variance differs, thus these are not observations from the same probability distribution.

Random sample

A proper **random sample** is i.i.d.

The law of large numbers and the Central Limit Theorem help us to predict the behaviour of our sample data.

Law of large numbers

The law of large numbers (LLN) states that, if these three assumptions are satisfied, the sample mean will approach the population mean with probability one if the sample is infinitely large.

Central Limit Theorem

If these three assumptions are satisfied,

Central Limit Theorem

If these three assumptions are satisfied,

- The sample mean is **normally distributed**, *regardless of the distribution of the original variable.*

Central Limit Theorem

If these three assumptions are satisfied,

- The sample mean is **normally distributed**, *regardless of the distribution of the original variable.*
- The sample mean has the **same expected value** as the population mean (LLN).

Central Limit Theorem

If these three assumptions are satisfied,

- The sample mean is **normally distributed**, *regardless of the distribution of the original variable.*
- The sample mean has the **same expected value** as the population mean (LLN).
- The standard deviation (**standard error**) of the sample mean is: $S.E.(\bar{x}) = \sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$.

Sample and population size

Note that the standard error depends only on the sample size, *not on the population size.*

Central Limit Theorem: unknown σ

When the population variance, σ, is unknown, we can use the sample estimate:

$$\hat{\sigma}_{\bar{x}} = \frac{\hat{\sigma}_x}{\sqrt{n}}$$
$$\hat{\sigma}_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Type I and II errors

Type I error: rejecting a null hypothesis that is true

Type I and II errors

Type I error: rejecting a null hypothesis that is true (e.g. $\alpha = .05$)

Type I and II errors

Type I error: rejecting a null hypothesis that is true (e.g. $\alpha = .05$)

Type II error: not rejecting a null hypothesis that is false

(Davidson & MacKinnon 1999: 126)

Power

Power: probability of rejecting a hypothesis that is false

$$1 - P(\text{Type II error})$$

Power

Power: probability of rejecting a hypothesis that is false

$$1 - P(\text{Type II error})$$

The power of a test increases when:

- the true value is further from the null hypothesis value;
- the variance is lower;
- the sample size is larger.

(Davidson & MacKinnon 1999: 126)

p-value

The p-value is the probability of a Type I error when rejecting the null hypothesis.

p-value

The p-value is the probability of a Type I error when rejecting the null hypothesis.

You can say a test is "statistically significant" if $p < \alpha$, but the p-value contains more information by itself.

(Davidson & MacKinnon 1999: 128)

$\alpha = .05$

Note that his value is absolutely arbitrary and just habit since the publication of Fisher (1923).

$\alpha = .05$

Note that his value is absolutely arbitrary and just habit since the publication of Fisher (1923).

The p-value is, one could argue, just a complicated way of measuring the sample size.

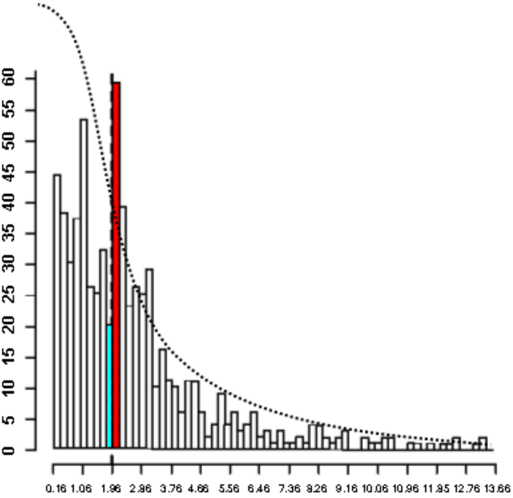
$\alpha = .05$

Note that his value is absolutely arbitrary and just habit since the publication of Fisher (1923).

The p-value is, one could argue, just a complicated way of measuring the sample size.

"Another interesting example (...) is the propensity for published studies to contain a disproportionally large number of Type I errors; studies with statistically significant results tend to get published, whereas those with insignificant results do not." (Kennedy 2008: 61)

$\alpha = .05$


0
5
10
15
20
25
30
35
40
45
50
55
60
0.16 1.06 1.96 2.86 3.76 4.66 5.56 6.46 7.36 8.26 9.16 10.06 10.96 11.85 12.76 13.66

Confidence intervals

Instead of an arbitrary threshold it is often more illuminating to present **confidence intervals** or graphical presentations of levels of uncertainty.

Outline

1. Probability distributions
2. Fundamentals
3. t- and F-tests

t-test

$$h_0 : \beta = 0$$
$$h_1 : \beta \neq 0$$

t-test

$$h_0 : \beta = 0$$
$$h_1 : \beta \neq 0$$

We can calculate the t-value by subtracting the value under the null and dividing by the standard error:

$$t = \frac{\hat{\beta}}{\sigma_{\hat{\beta}}}$$

t-test

$$h_0 : \beta = 0$$
$$h_1 : \beta \neq 0$$

We can calculate the t-value by subtracting the value under the null and dividing by the standard error:

$$t = \frac{\hat{\beta}}{\sigma_{\hat{\beta}}}$$

Because $\hat{\beta}$ has a normal distribution and $\sigma^2_{\hat{\beta}}$ a χ^2-distribution, t has the t-distribution with $n - k$ degrees of freedom.

t-test

```
bhat <- solve(t(x) %*% x) %*% t(x) %*% y
e <- y - x %*% bhat
vhat <- (1/(n-k) * t(e) %*% e) %x% solve(t(x) %*% x)
se <- sqrt(diag(vhat))
p <- 2 * (1 - pt(abs(bhat / se), n-k))
cbind(bhat, se, p)
```

`abs()`	absolute value
`2 *`	because it is a two-tailed test

t-test

$$h_0 : \beta = 0 \qquad h_1 : \beta \neq 0 \qquad t = \frac{\hat{\beta}}{\sigma_{\hat{\beta}}}$$

$$h_0 : \beta = a \qquad h_1 : \beta \neq a \qquad t = \frac{\hat{\beta} - a}{\sigma_{\hat{\beta}}}$$

$$h_0 : \beta_3 = \beta_2 \qquad h_1 : \beta_3 \neq \beta_2 \qquad t = \frac{\hat{\beta}_3 - \hat{\beta}_2}{\sigma_{\hat{\beta}_3}}$$

Sums of squares

SST Total sum of squares $\sum(y_i - \bar{y})^2$

SSE Explained sum of squares $\sum(\hat{y}_i - \bar{y})^2$

SSR Residual sum of squares $\sum e_i^2 = \sum(\hat{y}_i - y_i)^2 = \mathbf{e}'\mathbf{e}$

The key to remember is that **SST = SSE + SSR**

Sometimes instead of "explained" and "residual", "regression" and "error" are used, respectively, so that the abbreviations are swapped (!).

Anova

Variation	SS	df	MS
Explained	$\sum(\hat{y}_i - \bar{y})^2$	$k-1$	$SSE/(k-1)$
Residual	$\sum(\hat{y}_i - y_i)^2$	$n-k$	$SSR/(n-k)$
Total	$\sum(y_i - \bar{y})^2$	$n-1$	

Anova

Variation	SS	df	MS
Explained	$\sum(\hat{y}_i - \bar{y})^2$	$k-1$	$SSE/(k-1)$
Residual	$\sum(\hat{y}_i - y_i)^2$	$n-k$	$SSR/(n-k)$
Total	$\sum(y_i - \bar{y})^2$	$n-1$	

$$\frac{SSE/(k-1)}{SSR/(n-k)} \sim F_{k-1,n-k}$$

F-test

```
ssr <- t(e) %*% e
yhat <- X %*% bhat
sse <- t(yhat - mean(y)) %*% (yhat - mean(y))
F <- (sse/(k-1))/(ssr/(n-k))
1 - pf(F, k-1, n-k)
```

Restrictions

$$H_0 : \mathbf{y} = \mathbf{X}^{(1)}_{n \times k} \boldsymbol{\beta}_1 + \varepsilon$$
$$H_1 : \mathbf{y} = \mathbf{X}^{(1)}_{n \times k} \boldsymbol{\beta}_1 + \mathbf{X}^{(2)}_{n \times r} \boldsymbol{\beta}_2 + \varepsilon$$

USSR	Unrestricted sum of squared residuals
RSSR	Restricted sum of squared residuals

Restrictions

$$H_0 : \mathbf{y} = \mathbf{X}^{(1)}_{n \times k} \boldsymbol{\beta}_1 + \varepsilon$$
$$H_1 : \mathbf{y} = \mathbf{X}^{(1)}_{n \times k} \boldsymbol{\beta}_1 + \mathbf{X}^{(2)}_{n \times r} \boldsymbol{\beta}_2 + \varepsilon$$

USSR	Unrestricted sum of squared residuals
RSSR	Restricted sum of squared residuals

$$F_{\boldsymbol{\beta}_2} = \frac{(RSSR - USSR)/r}{USSR/(n-k-r)} \sim F(r, n-k-r)$$

Chow test

You might have a situation where you want to know whether the coefficients differ for two groups in the sample.

Chow test

You might have a situation where you want to know whether the coefficients differ for two groups in the sample.

You could test this with the following regression:

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \mathbf{0} \\ \mathbf{X}_2 \end{bmatrix} \gamma + \varepsilon,$$

whereby the difference between the coefficients for the two groups are captured by γ.

Chow test

It turns out you can just run two separate regressions ("unrestricted"), for the two groups, as well as ("restricted") regression $\mathbf{y}$ on $\mathbf{X}$, so that we get:

$$F_\gamma = \frac{(RSSR - SSR_1 - SSR_2)/k}{(SSR_1 + SSR_2)/(n - 2k)} \sim F(k, n - 2k)$$

(Davidson & MacKinnon 1999: 146-147)

Exercise: US wages

Open the `uswages.dta` data set and regress *log(wage)* on *educ*, *exper* and *race*.

1. Interpret all test results in the standard output.
2. Perform a test evaluating whether education and experience jointly contribute.
3. Perform a Chow test to see if the regression is different for urbanised versus rural respondents (*smsa*).
4. (bonus) Repeat the t- and F-tests using matrix algebra.

Appendix

Outline

Probability distributions in R

- You know x and want to know the **density** at that point: **d**

Probability distributions in R

- You know x and want to know the **density** at that point: **d**
- You know x and want to know the **area** up to that point: **p**

Probability distributions in R

- You know x and want to know the **density** at that point: **d**
- You know x and want to know the **area** up to that point: **p**
- You know x and want to know the area beyond that point: **1-p**

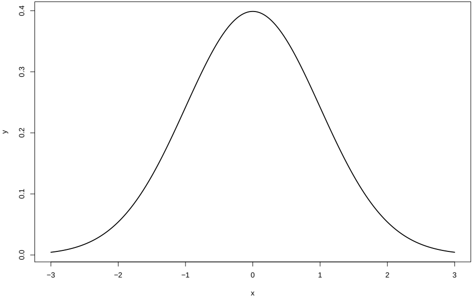
Probability distributions in R

- You know x and want to know the **density** at that point: **d**
- You know x and want to know the **area** up to that point: **p**
- You know x and want to know the area beyond that point: **1-p**
- You know the area and want to know the x value: **q**

Probability distributions in R

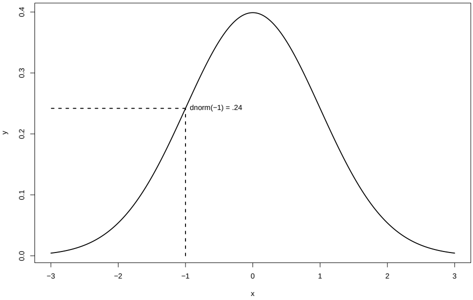
- You know x and want to know the **density** at that point: **d**
- You know x and want to know the **area** up to that point: **p**
- You know x and want to know the area beyond that point: **1-p**
- You know the area and want to know the x value: **q**
- You want **random** numbers drawn from that distribution: **r**

Probability distributions in R: dnorm



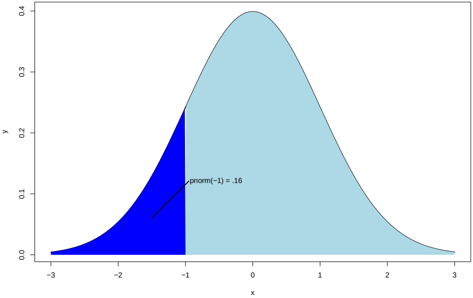

```
x <- seq(-3,3,.01)
y <- dnorm(x)
```

Probability distributions in R: dnorm



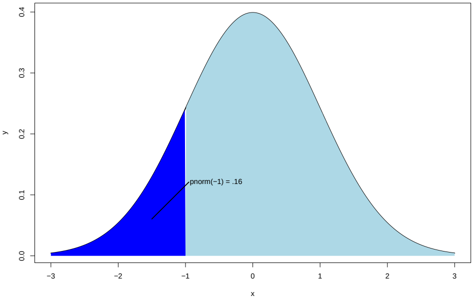

```
> dnorm(-1)
[1] 0.2419707
```

Probability distributions in R: pnorm



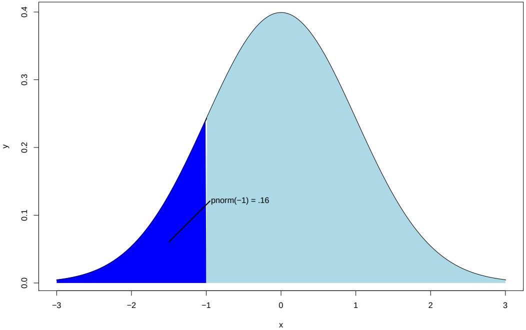

```
> pnorm(-1)
[1] 0.1586553
```

Probability distributions in R: pnorm



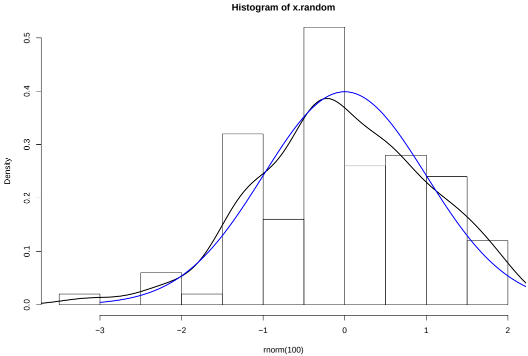

```
> 1 - pnorm(-1)
[1] 0.8413447
```

Probability distributions in R: pnorm



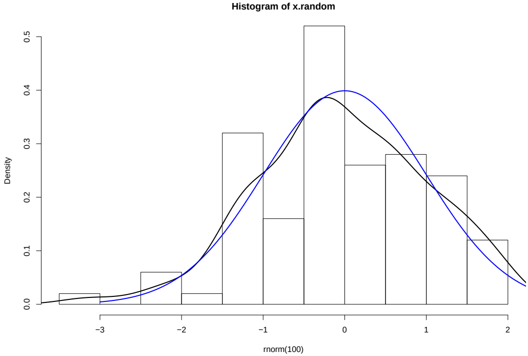

```
> qnorm(.1586553)
[1] -0.9999998
```

Probability distributions in R: rnorm




```
x.random <- rnorm(100)
```

Probability distributions in R: rnorm




```
hist(x.random, freq=false)
```

Probability distributions in R

These functions work on many distributions:

`rnorm()`	random from normal distribution
`pchisq()`	get area under χ^2-distribution
`1-pf()`	get area under *F*-distribution
`rbinom()`	draw randomly from binomial distribution
`1-dt()`	get *p*-value from *t*-distribution (one-tailed)

Probability distributions in R

Twenty throws (“Bernoulli trials”) with a coin:

```
> x <- rbinom(20,1,.5)
> factor(x, labels=c("head","tails"))
 [1] head  head  head  tails head  tails tails head
 [9] tails head  head  head  tails tails tails head
[17] tails tails tails head
```

Outline

Likelihood

"The likelihood that any parameter (or set of parameters) should have any assigned value (or set of values) is proportional to the probability that if this were so, the totality of observations should be that observed."

(Fisher 1922, 310)

Maximum Likelihood

The likelihood is proportional to the probability of observing the data you have (give or take some arbitrarily small deviation), given some parameter estimate:

$$L(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}) = \alpha P(\mathbf{y}, \mathbf{X}|\boldsymbol{\beta})$$

Maximum Likelihood

The likelihood is proportional to the probability of observing the data you have (give or take some arbitrarily small deviation), given some parameter estimate:

$$L(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}) = \alpha P(\mathbf{y}, \mathbf{X}|\boldsymbol{\beta})$$

The **likelihood function** is thus proportional to the **probability density function** of the given sample, as a function of the parameter values.

Maximum Likelihood

The likelihood is proportional to the probability of observing the data you have (give or take some arbitrarily small deviation), given some parameter estimate:

$$L(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}) = \alpha P(\mathbf{y}, \mathbf{X}|\beta)$$

The **likelihood function** is thus proportional to the **probability density function** of the given sample, as a function of the parameter values.

The estimator that maximizes this function also maximizes this probability density function and is the **Maximum Likelihood Estimator** (MLE or ML).

(Fisher 1922)

Three tests

Three tests are based on this likelihood:

- **Likelihood Ratio** (LR) test:

$$H_0 : 2 \log \frac{L_U}{L_R} = 2(\log L_U - \log L_R) = 0$$

Three tests

Three tests are based on this likelihood:

- **Likelihood Ratio** (LR) test:

$$H_0 : 2 \log \frac{L_U}{L_R} = 2(\log L_U - \log L_R) = 0$$

- **Wald** (W) test:

$$H_0 : \frac{(\hat{\beta}^{MLE} - \beta^*)^2}{var(\hat{\beta}^{MLE})} = 0$$

Three tests

Three tests are based on this likelihood:

- **Likelihood Ratio** (LR) test:

$$H_0 : 2 \log \frac{L_U}{L_R} = 2(\log L_U - \log L_R) = 0$$

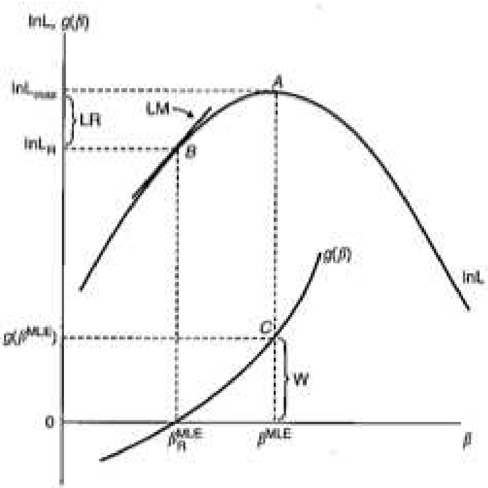
- **Wald** (W) test:

$$H_0 : \frac{(\hat{\beta}^{MLE} - \beta^*)^2}{var(\hat{\beta}^{MLE})} = 0$$

- **Langrange Multiplier** (LM) test:

$$H_0 : \frac{\partial lnL}{\partial \beta} = 0$$

Three tests



Three tests

Asymptotically, LR, W, and LM are all χ^2-distributed.

Three tests

Asymptotically, LR, W, and LM are all χ^2-distributed.

In small samples, $W \geq LR \geq LM$.

Three tests

Asymptotically, LR, W, and LM are all χ^2-distributed.

In small samples, $W \geq LR \geq LM$.

We return to these tests and how to do them in R later in the course.

(Kennedy 2008: 64)