Basic Statistics

Probability

Meaning

Prob(it will rain tomorrow) = 0.7

What does this mean? Surprisingly there is no single definition of this, there are at least 3 and maybe 5 or 6 that are incompatible and mutually contradictory.

Many fields, for example quantum mechanics (Copenhagen interpretation), have deep philosophical issues. In most fields these can be ignored for day to day work. Not in Statistics!

Conditional Probability $P(A|B)$

- P(Sum of 4 when rolling two fair dice) = 3/36
- P(Sum of 4 when rolling two fair dice | one die shows 5) = 0

Most important: $P(A|B) \neq P(B|A)$

- P(person is pregnant | person is female) $\approx$ 0.05
- P(person is female | person is pregnant) $\neq$ 0.05

Probability Distribution

models for outcomes of experiments.

Example we randomly select 10 undergraduates for a survey. What is the probability to get 5 male and 5 female students?

Each student is a *Bernoulli trial* with success probability 0.5. The sum of independent Bernoulli trials is *Binomial*, so

```r
round(dbinom(5, 10, 0.5), 3)
```

```
## [1] 0.246
```

Any statistical analysis begins by making an assumption about the probability distribution that describes the experiment.

Most famous line in Statistics: *we have observations from a normal distribution, ...*

If we are not sure we need to check! (goodness-of-fit testing)

These models always include unknown numbers (parameters). A large part of Statistics is to use data to estimate those numbers.

Probability vs Statistics

- Probability let's us make a guess about what outcomes will be like. (1 in 5 chance of equal numbers of male and female students)
- Statistics let's use use data to guess (estimate) what the exact distribution is like.

Frequentist vs Bayesian Statistics

two ways to do everything.

they already differ in the meaning of the word *probability*.

Essential difference:

- Frequentist Statistics: P(Data|Theory)
- Bayesian Statistics: P(Theory|Data)

everyone agrees: P(Theory | Data) is what we want.

Example P(Higgs boson exists | data from CMS and Atlas)

BUT

Bayes formula:

P(Theory|Data) = P(Data|Theory)P(Theory)/P(Data)

- P(Theory) *prior distribution*

Example what was P(Higgs boson exists) in (say) 2000?

- P(Data) multi-dimensional integral, often impossible to find even with numerical methods.

With advances like *Markov Chain Monte Carlo* methods Bayesian statistics has become much more widely used in the last 10 to 15 years.

Confidence Intervals

In general when we estimate a parameter we also want an idea of the *error* in that estimate. For this we can find a *confidence interval*.

Example In a random sample of 50 physicists 34 say that super-symmetry is real. What is a 90% confidence interval for the true proportion among all physicists that believe that?

Each physicist is a Bernoulli trial with a success probability p=P(believes super-symmetry is real). So the number who say so in our sample is binomial with n=50 and p. Now

```r
round(as.numeric(binom.test(x=34, n=50, conf.level = 0.9)$conf.int), 3)
```

```
## [1] 0.555 0.788
```

What exactly does this mean? It is that over the long run with many people calculating 90% confidence intervals, 90% of those intervals will be “good” (aka actually include the interval) and 10% will not.

A simple simulation:

```r
x <- rbinom(100, 50, 0.65) # true percentage is 65
df <- data.frame(x=x, L=0*x, R=0*x,
         Good=rep("BAD!!!", 100))
for(i in 1:100) {
  ci <- round(as.numeric(binom.test(x=x[i], n=50,
              conf.level = 0.9)$conf.int), 3)
  df$L[i] <- ci[1]
  df$R[i] <- ci[2]
  if(ci[1]<0.65 & 0.65<ci[2]) df$Good[i] <- "Good"
}
kable.nice(df)
```

x	L	R	Good
31	0.494	0.735	Good
35	0.576	0.805	Good
34	0.555	0.788	Good
34	0.555	0.788	Good
34	0.555	0.788	Good
33	0.535	0.770	Good
34	0.555	0.788	Good
29	0.454	0.699	Good
36	0.597	0.822	Good
34	0.555	0.788	Good
33	0.535	0.770	Good
37	0.619	0.839	Good
25	0.376	0.624	BAD!!!
29	0.454	0.699	Good
34	0.555	0.788	Good
34	0.555	0.788	Good
33	0.535	0.770	Good
35	0.576	0.805	Good
32	0.514	0.753	Good
27	0.415	0.662	Good
34	0.555	0.788	Good
32	0.514	0.753	Good
34	0.555	0.788	Good
33	0.535	0.770	Good
34	0.555	0.788	Good
31	0.494	0.735	Good
33	0.535	0.770	Good
30	0.474	0.717	Good
33	0.535	0.770	Good
35	0.576	0.805	Good
34	0.555	0.788	Good
36	0.597	0.822	Good
32	0.514	0.753	Good
32	0.514	0.753	Good
34	0.555	0.788	Good
38	0.640	0.855	Good
34	0.555	0.788	Good
33	0.535	0.770	Good
28	0.434	0.680	Good
34	0.555	0.788	Good
25	0.376	0.624	BAD!!!

so in the long run 90% of the intervals are "Good", that is include the true percentage.

But is our interval (0.555, 0.788) a good one? Nobody knows!

Coverage

The basic property of a confidence interval is its *coverage*, that is the actual probability that the interval includes the true parameter.

Many (most?) methods are based on some approximations (central limit theorem etc.), so coverage is not guaranteed and needs to be checked.

Example coverage of standard t intervals for a population mean:

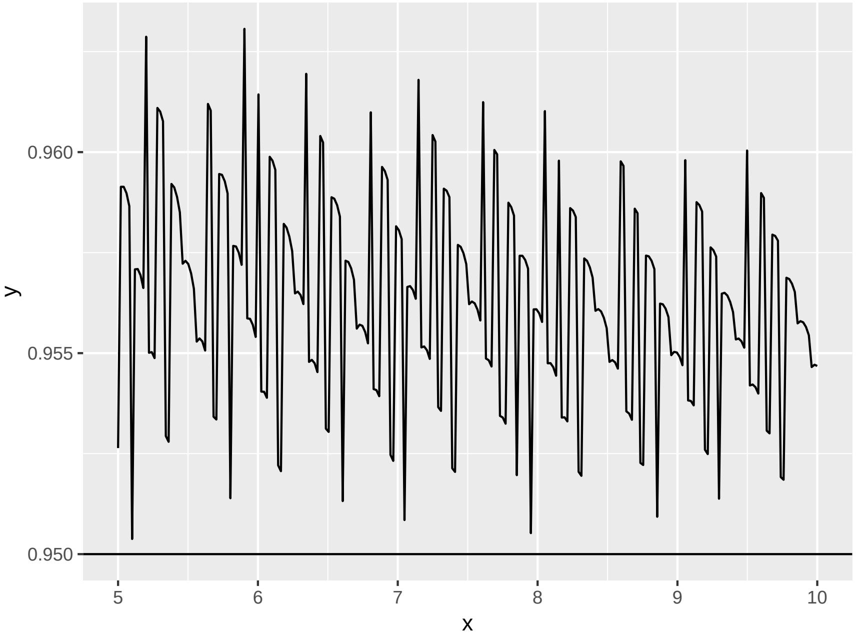
```
mu <- 0; sigma <- 1; n <- 50
k <- 0
for(i in 1:1000) {
  x <- rnorm(n, mu, sigma) #create data
  ci <- t.test(x)$conf.int #find interval
  if(ci[1]<0 & 0<ci[2]) k <- k+1 #check
}
k/1000
```

```
## [1] 0.949
```

now repeat for all (many) values of mu, sigma and n. Usually shown as a graph.

Sometimes these look strange:

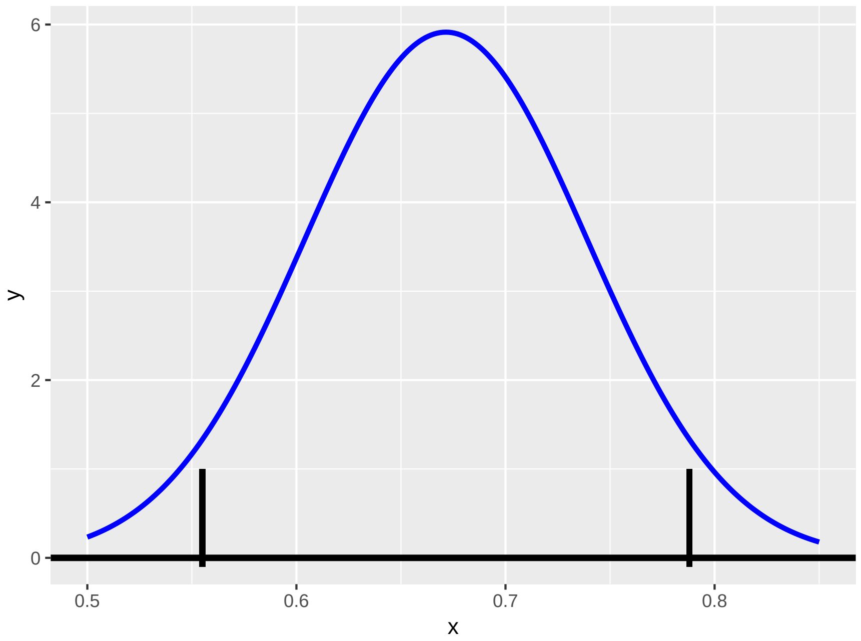
Example Garwood intervals for Poisson rate



This is because the Poisson is a discrete random variable.

Problems with Confidence Intervals

There is a common misinterpretation of confidence intervals. People often have this mental image:



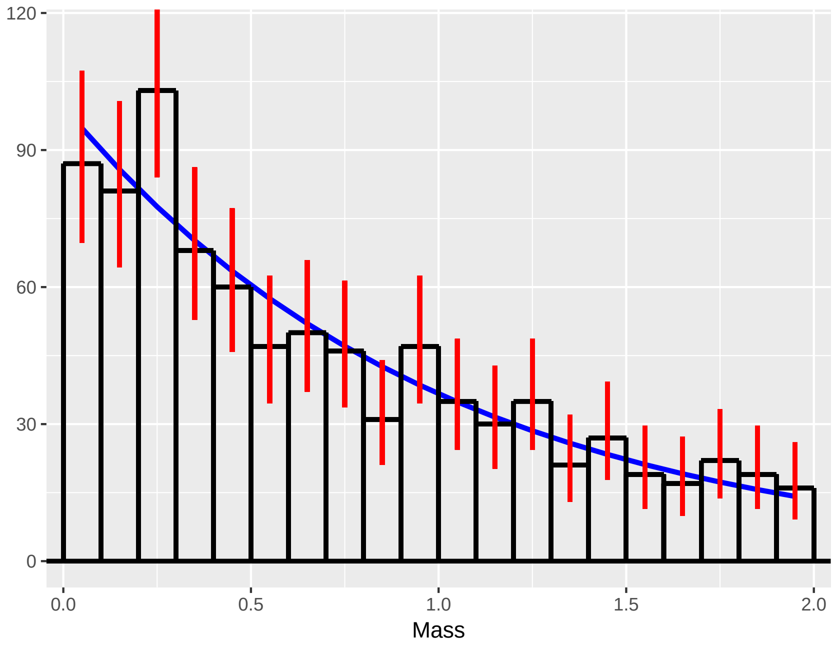
this would suggest that a value in the middle (say 0.7) is more likely than a value on the edge (say 0.75)

This is wrong!

Actually, it is the correct interpretation if the interval is a *Bayesian credible interval*.

Another typical issue with confidence intervals:

Here is a classic graph in physics:



The idea is this: we have a model (blue curve) and we have a histogram of data. Each red horizontal line segment is a 95% confidence interval for the bin counts. If the model is good only a few of these should not include the curve.

Problem? The intervals are *point-wise*, so the 95% applies to each individually. As above if there many some (95%!) will not include the correct value by random fluctuation.

This is an example of **simultaneous inference**, a very common and hard problem is Statistics.

One solution: Bonferroni. Say intervals are independent, then

$$\begin{aligned}
&w = P(\text{at least one interval bad}) = \\
&1 - P(\text{all intervals good}) = \\
&1 - \prod_{i=1}^{k} P(i^{th} \text{ interval good}) = \\
&1 - \prod_{i=1}^{k} (1-\alpha) = 1 - (1-\alpha)^k
\end{aligned}$$

so if we find $1-(1-\alpha)^{1/k}$ intervals, the collection will have probability of $1-\alpha$ to be good.

But our intervals are not independent because neighboring bins have similar counts!

Big problem in ML!!

Hypothesis Testing

methods designed to answer yes-no questions.

Example Does super-symmetry exist?

in some ways much harder than confidence intervals, in other ways they are the same thing.

Needs:

- null hypothesis
- alternative hypothesis
- type I error probability $\alpha = \text{P}(\text{reject } H_0 \mid H_0 \text{ true})$
- p-value

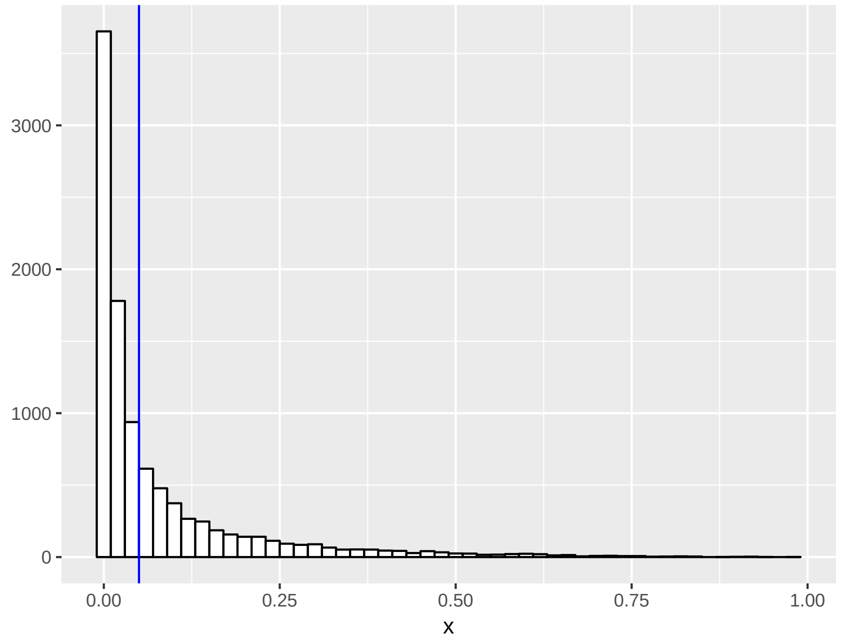
also important to consider the power of the test, that is $\text{P}(\text{reject } H_0 \mid H_0 \text{ false})$.

Many issues and misunderstandings:

- failing to reject the null is NOT the same as accepting the null.
- p value is NOT the probability that the null hypothesis is true.
- if null is false it will always be rejected as long as the sample size is large enough.
- Dichotomizing the p value with $p < \alpha \rightarrow$ reject null

Example We have data from normal distribution with standard deviation 1. We test $H_0 : \mu = 0$ vs $H_a : \mu > 0$. But really $\mu = 2$. Now:

```r
B <- 10000; mu <- 2
x <- matrix(rnorm(B, mu), nrow=B)
TS <- apply(x, 1, mean) #test statistic(s)
df <- data.frame(pvals = 1-pnorm(TS))
ggplot(df, aes(pvals)) +
  geom_histogram(,
    color = "black",
    fill = "white",
    binwidth = 1/50) +
    labs(x = "x", y = "") +
  geom_vline(xintercept = 0.05, color="blue")
```



Issues are serious enough that in some fields they are banning the use of hypothesis tests. To Statisticians this is an over-reaction, but the problems with the practical use of hypothesis testing are real!!

Statistics has many facets:

- Non parametric methods
- ANOVA
- Ordinary Regression
- Transformations and polynomial regression
- non parametric regression
- logistic regression
- generalized additive models
- general linear models
- principal components
- ridge regression and the lasso
- regression trees

- non parametric density estimation
- survival analysis
- quality control, six sigma
- time series analysis
- goodness of fit
- classification with LDA, QDA, Trees, Boosting, Support vector machines, neural networks, deep learning

. . .

Many (all?) have connections to machine learning

Further Reading

- Standard Textbook for Probability and Statistics: Casella and Berger: Statistical Inference
- Any of my courses at http://academic.uprm.edu/wrolke/
- Issues with hypothesis testing: What's Wrong with Hypothesis Testing