### More Useful Statistical Models

**The normal (or Gaussian) distribution**

Generating process: An estimate with uncertainty; a measurement with error.

Parameters:
- mean: the most likely value
- standard deviation: “the spread”

Getting random values:
- Excel: `=NORM.INV(RAND(),mean,stderr)`
- R: `rnorm(N,mean,stderr)`

![Normal distribution](image)

### The gamma distribution

Generating process: Website latency; a sequence of occasionally (exponentially) slow tasks

Parameters:
- scale: the average time to complete a subtask
- n: the number of subtasks

Getting random values:
- Excel: `=GAMMA.INV(RAND(),n,scale)`
- R: `rgamma(N,n,1/scale)`

![Gamma distribution](image)

---

### Decisions in the Face of Uncertainty - Just Enough Statistics to be Dangerous

Presenter: John Rauser
O’Reilly Velocity Conference - Santa Clara, CA
June 14, 2011

**Estimation Questions (Don’t use the Internet!)**

<table>
<thead>
<tr>
<th>Question</th>
<th>90% interval</th>
<th>Correct?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) How many passengers flew into or out of San José International Airport (SJC) in 2009?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2) According to the May 26, 2011 run of The HTTP Archive, what percentage of sites used jQuery?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3) What is the mass of the planet Venus, expressed as a multiple of Earth’s mass?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4) How many bones are there in a normal adult human skeleton?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5) The people of the Czech Republic consume more beer per capita than any other nation. How much beer does the average Czech drink per year (in liters or pints)?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6) How many inches of rain fell on Seattle during the month of December 2010?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7) According to the May 26, 2011 run of The HTTP Archive, how big was the average page in kB, including all assets?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8) Is the US, what percentage of babies born are girls?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9) @ladygaga is the most followed user on twitter. How many followers did she have on May 31, 2011?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10) How many elements (e.g. <code>&lt;p&gt;</code>, <code>&lt;div&gt;</code>, etc.) are defined by the HTML 4.01 specification?</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Exploring a binomial model in R

Simulating one calibrated estimator taking a ten question test. How many did she get right?

```r
> rbinom(1,10,0.9)
[1] 6
```

Simulating fifteen calibrated estimators each taking a ten question test. How many did each one get right?

```r
> rbinom(15,10,0.9)
[1] 10 10 10  7  9  9  9  8 10  9  9  8 10 10
```

Finding the average number correct among a simulated class of fifteen.

```r
> mean(rbinom(15,10,0.9))
```

Plotting the results from 10,000 simulated classes of 20 students.

```r
> library(ggplot2)
> qplot(replicate(10000, mean(rbinom(20, 10, 0.9))), geom="histogram", binwidth=1/20)
```

You should treat with great suspicion a group of 20 that averages more than 9.5 or less than 8.5. These are rare occurrences. In 10,000 simulated classes we never saw a class of 20 average less than 7.9 on 10 questions.

Useful Statistical Models

The Bernoulli distribution

Generating process: A coin toss, a single success/fail trial

Parameters:
- \( p \): the probability of success

Getting random values:
- Excel: \( =\text{IF}(\text{RAND()} > p,"S","F") \)
- R: \( \text{sample(c("S","F"),N,replace=TRUE,prob=c(p,1-p))} \)

The binomial distribution

Generating process: A sequence of success/fail trials

Parameters:
- \( p \): the probability of success
- \( n \): the number of trials

Getting random values:
- Excel: \( =\text{BINOM.INV}(n,p,\text{RAND()}) \)
- R: \( \text{rbinom(N,n,p)} \)