Advanced Econometric Methods EMET3011/8014

Lecture 5

John Stachurski

Semester 1, 2011

Notes

Errata:

- k-th moment of x is $\mathbb{E}\left[x^k\right]$ not $\mathbb{E}\left[|x|^k\right]$
- Further typos/corrections on course homepage

Assessment:

- Mid semester exam: May 5 — see course homepage

Today's Lecture

- Finish probability theory
- Move on to statistics

Best Linear Predictors

Let x and y be RVs, where x is observed before y

Suppose we want to predict y given x

Mathematically: find a function $g \colon \mathbb{R} \to \mathbb{R}$ such that

$$g(x) \text{ is "close" to } y \text{ "on average"}$$

Closeness measured by mean squared deviation, so problem is

$$\min_{g \in \mathscr{G}} \mathbb{E}\left[(y - g(x))^2\right] \quad \text{where} \quad \mathscr{G} := \{\text{all } g \colon \mathbb{R} \to \mathbb{R}\}$$

Minimizer is $g^*(x) := \mathbb{E}[y \mid x]$—we'll discuss it later

Simplification: Predict y with an <u>affine</u> function of x

Set of all affine functions:

$$\mathcal{L} := \{ \text{ all functions of the form } g(x) = \alpha + \beta x \}$$

Thus, we consider the problem $\min_{g \in \mathcal{L}} \mathbb{E}\left[(y - g(x))^2\right]$

Equivalent problem: $\min_{\alpha, \beta \in \mathbb{R}} \mathbb{E}\left[(y - \alpha - \beta x)^2\right]$

Objective function can be written as

$$\mathbb{E}\left[y^2\right] - 2\alpha\mathbb{E}\left[y\right] - 2\beta\mathbb{E}\left[xy\right] + 2\alpha\beta\mathbb{E}\left[x\right] + \alpha^2 + \beta^2\mathbb{E}\left[x^2\right]$$

Take derivatives, set equal to zero, solve simultaneously

Exercise: Show that the minimizers are

$$\beta^* := \frac{\mathrm{cov}[x, y]}{\mathrm{var}[x]} \quad \text{and} \quad \alpha^* := \mathbb{E}[y] - \beta^* \mathbb{E}[x]$$

Best linear predictor is then

$$\ell^*(x) := \alpha^* + \beta^* x$$

Exercise: Show that $\mathbb{E}[\ell^*(x)] = \mathbb{E}[y]$

Exercise: Compare α^*, β^* with expressions for estimated coefficients in simple OLS. Can you see some similarity?

Common Distributions

Let's list some well-known distributions needed for this course

The **uniform distribution** on $[a, b] \subset \mathbb{R}$ has density

$$p(s; a, b) := \frac{1}{b-a} \mathbb{1}\{a \leq s \leq b\}$$

- Represent symbolically by $U[a, b]$
- Exercise: Show that mean $= (a+b)/2$

The univariate **normal** or **Gaussian distribution** has density

$$p(s; \mu, \sigma) := (2\pi\sigma^2)^{-1/2} \exp\left\{ -\frac{(s-\mu)^2}{2\sigma^2} \right\}$$

Comments:

- $\mu \in \mathbb{R}$ and $\sigma > 0$
- Represented symbolically by $\mathcal{N}(\mu, \sigma^2)$
- $\mathcal{N}(0, 1)$ is called the **standard normal distribution**

Facts:

- If $x \sim \mathcal{N}(\mu, \sigma^2)$, then $\mathbb{E}[x] = \mu$, and $\mathrm{var}[x] = \sigma^2$
- If $x_n \sim$ normal and $\alpha_n \in \mathbb{R}$, then $\alpha_0 + \sum_{n=1}^{N} \alpha_n x_n \sim$ normal

The **chi-squared distribution with k degrees of freedom** has density

$$p(s;k) := \frac{1}{2^{k/2}\Gamma(k/2)} s^{k/2-1} e^{-s/2} \qquad (s \geq 0)$$

- Γ is the gamma function — def omitted
- Represented symbolically by $\chi^2(k)$

Facts:

- If $x_1, \ldots, x_k \overset{\text{IID}}{\sim} \mathcal{N}(0,1)$, then $\sum_{i=1}^k x_i^2 \sim \chi^2(k)$
- If $Q_j \sim \chi^2(k_j)$, independent, then $\sum_{j=1}^J Q_j \sim \chi^2(\sum_{j=1}^J k_j)$

The **Student's t-distribution with k degrees of freedom** has density

$$p(s;k) := \frac{\Gamma(\frac{k+1}{2})}{(k\pi)^{1/2}\Gamma(\frac{k}{2})}\left(1+\frac{s^2}{k}\right)^{-(k+1)/2}$$

Fact: If

- $Z \sim \mathcal{N}(0,1)$,
- $Q \sim \chi^2(k)$, and
- Z and Q are independent,

then $Z(k/Q)^{1/2} \sim$ t-distribution with k df

F-distribution

The **F-distribution** with parameters k_1, k_2 has density

$$p(s; k_1, k_2) := \frac{\sqrt{(k_1 s)^{k_1} k_2^{k_2} / [k_1 s + k_2^{k_1 + k_2}]}}{s B(k_1/2, k_2/2)} \qquad (s \geq 0)$$

- B is the Beta function — def omitted
- Represented symbolically by $F(k_1, k_2)$

Fact: If $Q_1 \sim \chi^2(k_1)$ and $Q_2 \sim \chi^2(k_2)$ are independent, then

$$\frac{Q_1/k_1}{Q_2/k_2} \sim F(k_1, k_2)$$

Asymptotics

Common statistical problem:

- How do different estimators, tests, etc. perform as amount of data goes to ∞?

To this end, we now investigate asymptotic theory

(Distributions of limits of sequences of random variables)

Main tools: LLN and CLT

Convergence in Probability

Let $\{x_n\}_{n=1}^{\infty}$ be a sequence of RVs, x another RV

Def: $\{x_n\}_{n=1}^{\infty}$ converges to x **in probability** ($x_n \xrightarrow{p} x$) if,

$$\text{given any } \delta > 0, \quad \mathbb{P}\{|x_n - x| > \delta\} \to 0 \quad \text{as } n \to \infty$$

Often limit is constant

Example: Claim that if $x_n \sim \mathcal{N}(\alpha, 1/n)$, then $x_n \xrightarrow{p} \alpha$

Here's the picture (formal proof later)

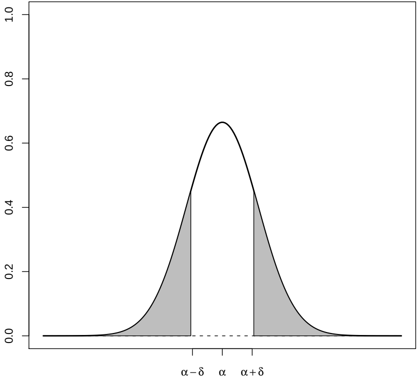


Figure: $\mathbb{P}\{|x_n - \alpha| > \delta\} \to 0$

Here's the picture (formal proof later)

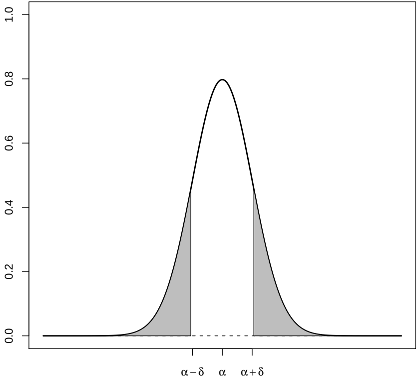


Figure: $\mathbb{P}\{|x_n - \alpha| > \delta\} \to 0$

Here's the picture (formal proof later)

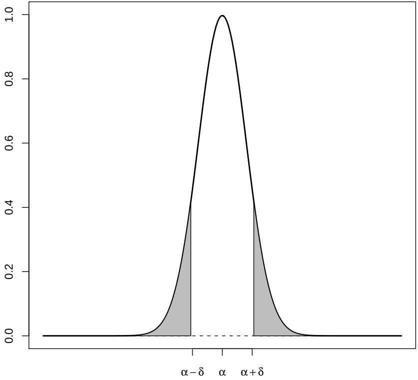


Figure: $\mathbb{P}\{|x_n - \alpha| > \delta\} \to 0$

Following statements are true:

1. If $g\colon \mathbb{R} \to \mathbb{R}$ is continuous and $x_n \xrightarrow{p} x$, then $g(x_n) \xrightarrow{p} g(x)$
2. If $x_n \xrightarrow{p} x$ and $y_n \xrightarrow{p} y$, then

$$x_n + y_n \xrightarrow{p} x + y \quad \text{and} \quad x_n y_n \xrightarrow{p} xy$$

3. If $x_n \xrightarrow{p} x$ and $\alpha_n \to \alpha$, then

$$x_n + \alpha_n \xrightarrow{p} x + \alpha \quad \text{and} \quad x_n \alpha_n \xrightarrow{p} x\alpha$$

- Here $\{\alpha_n\}$ is a nonrandom scalar sequence

Convergence in Mean Square

Def: $\{x_n\}_{n=1}^{\infty}$ converges to x **in mean square** ($x_n \stackrel{ms}{\to} x$) if

$$\mathbb{E}\left[(x_n - x)^2\right] \to 0 \quad \text{as } n \to \infty$$

Fact: If $x_n \stackrel{ms}{\to} x$, then $x_n \stackrel{p}{\to} x$

Follows from **Chebychev's inequality**: Given RV z,

$$\mathbb{P}\{|z| \geq \delta\} \leq \frac{\mathbb{E}\left[z^2\right]}{\delta^2} \quad \text{for any } \delta > 0$$

$$\therefore \quad 0 \leq \mathbb{P}\{|x_n - x| > \delta\} \leq \mathbb{P}\{|x_n - x| \geq \delta\} \leq \frac{\mathbb{E}\left[(x_n - x)^2\right]}{\delta^2}$$

Fact: If α is constant, then $x_n \stackrel{ms}{\to} \alpha$ whenever

1. $\mathbb{E}[x_n] \to \alpha$
2. $\mathrm{var}[x_n] \to 0$

True because $\mathbb{E}[(x_n - \alpha)^2] = \mathrm{var}[x_n] + (\mathbb{E}[x_n] - \alpha)^2$

- Exercise: Verify this equality

Example: If $x_n \sim \mathcal{N}(\alpha, 1/n)$, then $x_n \stackrel{p}{\to} \alpha$

Proof:

$$\mathbb{E}[x_n] = \alpha \to \alpha \quad \text{and} \quad \mathrm{var}[x_n] = 1/n \to 0$$

Convergence in distribution

Let $\{F_n\}_{n=1}^{\infty}$ be a sequence of cdfs, and let F be a cdf

Def: $\{F_n\}_{n=1}^{\infty}$ **converges weakly** to F if, for any s such that F is continuous at s, we have

$$F_n(s) \to F(s) \quad \text{as } n \to \infty$$

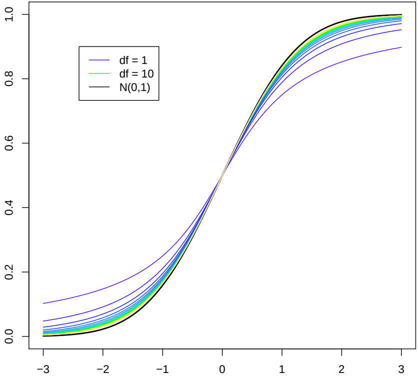


Figure: t-distribution with k df converges to $\mathcal{N}(0,1)$ as $k \to \infty$

Def: $\{x_n\}_{n=1}^{\infty}$ converges to x **in distribution** ($x_n \xrightarrow{d} x$) if

$$x_n \sim F_n, \quad x \sim F \quad \text{and} \quad F_n \to F \text{ weakly}$$

Following statements are true:

1. If $g \colon \mathbb{R} \to \mathbb{R}$ is continuous and $x_n \stackrel{d}{\to} x$, then $g(x_n) \stackrel{d}{\to} g(x)$
2. If $x_n \stackrel{p}{\to} x$, then $x_n \stackrel{d}{\to} x$
3. If α is constant, $x_n \stackrel{p}{\to} \alpha$ and $y_n \stackrel{d}{\to} y$, then

$$x_n + y_n \stackrel{d}{\to} \alpha + y \quad \text{and} \quad x_n y_n \stackrel{d}{\to} \alpha y$$

- Fact 1 called the **continuous mapping theorem**
- Fact 3 called **Slutsky's theorem**

LLN and CLT

LLN = law of large numbers

CLT = central limit theorem

Two of most important theorems in statistics

LLN: sample means converge to means (i.e., expectations)

CLT: Averages are asymptotically normal

Law of Large Numbers

Let $\{x_n\}_{n=1}^{\infty} \stackrel{\text{IID}}{\sim} F$ and $\bar{x}_N := \frac{1}{N}\sum_{n=1}^{N} x_n$

Theorem. If $\int |s| F(ds) < \infty$, then

$$\bar{x}_N \xrightarrow{p} \mathbb{E}\left[x_n\right] = \int sF(ds) \quad \text{as} \quad N \to \infty$$

The proof is an important exercise—use facts we've discussed

See corresponding exercise in the course notes (solution provided)

Illustration of LLN with R:

Consider flipping a coin until 10 heads have occurred

Probability of heads is 0.4

Let x be number of tails observed in the process

Can show analytically that mean $\mathbb{E}[x]$ is 15

Let's check LLN with a simulation

```r
N <- 10000
outcomes <- numeric(N)
for (i in 1:N) {
    num.tails <- 0
    num.heads <- 0
    while (num.heads < 10) {
        b <- runif(1)
        num.heads <- num.heads + (b < 0.4)
        num.tails <- num.tails + (b >= 0.4)
    }
    outcomes[i] <- num.tails
}
print(mean(outcomes))
```

Running program gives values close to 15

Second version:

```
# Define function to simulate draws of x
# Parameter q is probability of heads in each flip
f <- function(q) {
    num.tails <- 0
    num.heads <- 0
    while (num.heads < 10) {
        b <- runif(1)
        num.heads <- num.heads + (b < q)
        num.tails <- num.tails + (b >= q)
    }
    return(num.tails)
}

# Generate 10^5 observations of x, print sample mean
outcomes <- replicate(10000, f(0.4))
print(mean(outcomes))
```

LLN more general than it looks:

If $\{x_n\}_{n=1}^{\infty} \overset{\text{IID}}{\sim} F$ and $h \colon \mathbb{R} \to \mathbb{R}$ with $\int |h(s)| F(ds) < \infty$, then

$$\frac{1}{N} \sum_{n=1}^{N} h(x_n) \xrightarrow{p} \mathbb{E}\,[h(x_n)] := \int h(s) F(ds) \tag{1}$$

Proof: If $y_n := h(x_n)$ LLN gives $\frac{1}{N} \sum_{n=1}^{N} y_n \xrightarrow{p} \mathbb{E}\,[y_n]$ which is (1)

Example: set $h(s) = s^2$ to get

$$\frac{1}{N} \sum_{n=1}^{N} x_n^2 \xrightarrow{p} \mathbb{E}\,[x_n^2] \quad \text{as} \quad N \to \infty$$

LLN applies to probabilities as well

Claim: For any $B \subset \mathbb{R}$,

$$\frac{1}{N} \sum_{n=1}^{N} \mathbb{1}\{x_n \in B\} \xrightarrow{p} \mathbb{P}\{x_n \in B\}$$

Proof: Let $h(s) := \mathbb{1}\{s \in B\}$

Expectations of indicators equal probabilities of events, so

$$\mathbb{E}\left[h(x_n)\right] = \mathbb{E}\left[\mathbb{1}\{x_n \in B\}\right] = \mathbb{P}\{x_n \in B\}$$

Therefore, by LLN,

$$\frac{1}{N} \sum_{n=1}^{N} \mathbb{1}\{x_n \in B\} = \frac{1}{N} \sum_{n=1}^{N} h(x_n) \xrightarrow{p} \mathbb{E}\left[h(x_n)\right] = \mathbb{P}\{x_n \in B\}$$

Central Limit Theorem

Let $\{x_n\}_{n=1}^{\infty} \overset{\text{IID}}{\sim} F$ with

- $\mu := \mathbb{E}[x_n] = \int sF(ds)$
- $\sigma^2 := \text{var}[x_n] = \int (s-\mu)^2 F(ds)$

Theorem: If $\int s^2 F(ds) < \infty$, then

$$\sqrt{N}(\bar{x}_N - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2) \quad \text{as} \quad N \to \infty$$

Proof: Omitted

Follows that

$$\sqrt{N}\left\{\frac{\bar{x}_N - \mu}{\sigma}\right\} \xrightarrow{d} \mathcal{N}(0,1)$$

Proof: If $y \sim \mathcal{N}(0, \sigma^2)$, then

$$\sqrt{N}(\bar{x}_N - \mu) \xrightarrow{d} y$$

Applying continuous mapping theorem, we get

$$\sqrt{N}\left\{\frac{\bar{x}_N - \mu}{\sigma}\right\} \xrightarrow{d} \frac{y}{\sigma}$$

Clearly y/σ is normal,

$$\mathbb{E}\left[y/\sigma\right] = 0 \quad \text{and} \quad \operatorname{var}\left[\frac{y}{\sigma}\right] = \frac{1}{\sigma^2}\operatorname{var}[y] = \frac{1}{\sigma^2}\sigma^2 = 1$$

CLT tells us about distribution of $\bar{x}_N$ when N large

Informally, for N large we have

$$\sqrt{N}(\bar{x}_N - \mu) \approx y \sim \mathcal{N}(0, \sigma^2)$$

Therefore,

$$\bar{x}_N \approx \frac{y}{\sqrt{N}} + \mu \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{N}\right)$$

Thus, $\bar{x}_N$ approximately normally distributed, with

- mean equal to μ, and
- variance $\to 0$ at rate proportional to $1/N$

Simulation Exercise

Let's illustrate the convergence in

$$z_N := \sqrt{N}\left\{\frac{\bar{x}_N - \mu}{\sigma}\right\} \xrightarrow{d} \mathcal{N}(0,1)$$

Let $x_n \sim \chi^2(5)$ — an arbitrary choice

Note:

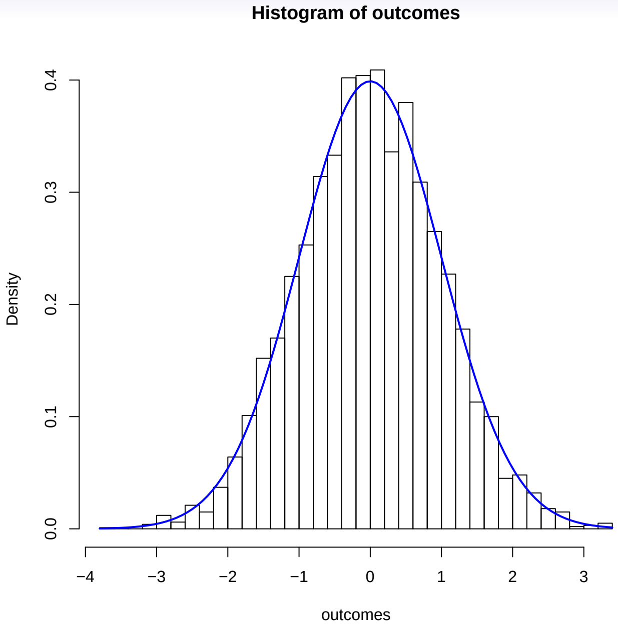
- $\mathbb{E}\left[x_n\right] = 5$
- $\mathrm{var}[x_n] = 2 \times 5 = 10$

Next listing generates 5,000 observations of z_N when $N = 10^3$

```
num.replications <- 5000
obs <- numeric(num.replications)
N <- 1000
k <- 5      # Degrees of freedom

for (i in 1:num.replications) {
    xvec <- rchisq(N, k)
    obs[i] <- sqrt(N / (2 * k)) * (mean(xvec) - k)
}

hist(obs, breaks=50, freq=FALSE)
curve(dnorm, add=TRUE, lw=2, col="blue")
```


Histogram of outcomes
Density
0.0
0.1
0.2
0.3
0.4
−4
−3
−2
−1
0
1
2
3
outcomes

Exercise: Experiment with different distributions for x_n

- binomial
- F
- exponential
- poisson, etc

Should still get good fit to $\mathcal{N}(0, 1)$ whenever second moment finite

Extension to the CLT:

Let

- $\{x_n\}_{n=1}^{\infty}$ be as in CLT
- $g\colon \mathbb{R} \to \mathbb{R}$ be differentiable at $\mu = \mathbb{E}[x_n]$

Theorem: If $g'(\mu) \neq 0$, then

$$\sqrt{N}\{g(\bar{x}_N) - g(\mu)\} \xrightarrow{d} \mathcal{N}(0, g'(\mu)^2\sigma^2) \quad \text{as} \quad N \to \infty$$

Used frequently in statistics to obtain asymptotic distributions

The technique is referred to as the **delta method**

Statistical Learning

Now we switch from probability theory to statistics

What's the difference?

- Probability: Try to guess outcomes from probabilities
- Statistics: Try to guess probabilities from outcomes

Generalization

The fundamental problem of statistics: Learning from data

Learning from data is concerns generalization

Example: A certain drug tested on 1,000 volunteers

- Found to produce the desired effect in 95% of cases
- Drug company now claims drug is highly effective
- Underlying assertion: Can generalize to the wider population
- Outcome for volunteers has implications for other people

Another word for generalization is induction

Inductive learning: Reasoning proceeds from specific to general

1. You show a child pictures of dogs in a book and say 'dog'
2. The child sees a dog on the street and says 'dog'

Deductive learning: Reasoning proceeds from general to specific

1. You tell a child that dogs are hairy, four legged animals that stick their tongues out when hot
2. Child determines animal is a dog on this basis

Statistical learning inductive, not deductive

Typical statistical problems:

Example

N random values $x_1, \ldots, x_N$ are drawn from a fixed but unknown cdf F. We wish to learn about F from this sample

Example

Same as above, but now we only care about learning the mean of F—or the standard deviation, or the median, etc.

Unknown quantities/functions must be inferred from the sample

Example

Observe

- "inputs" $x_1, \ldots, x_N$ to some "system"
- corresponding "outputs" $y_1, \ldots, y_N$

Problem: Find f such that $f(x)$ will be a good guess of y

If "good guess" means minimal mean squared error, then best choice of f is $f(x) = \mathbb{E}[y \mid x]$

But <u>we do not know the underlying distributions</u>

- Hence cannot compute $\mathbb{E}[y \mid x]$
- Must do our best from info contained in sample

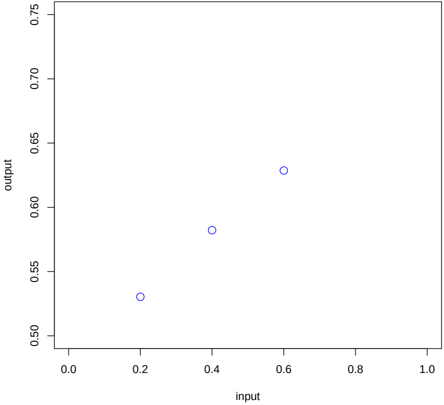
Assumptions and Learning

Most learning/generalization requires more than just data

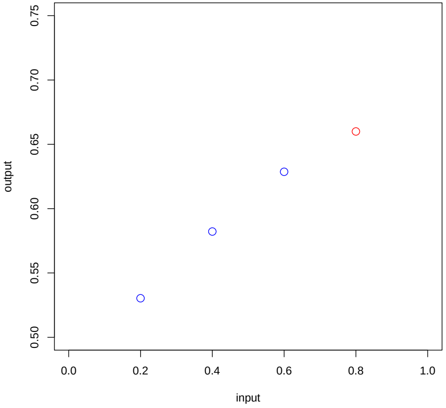
Example: Consider following problem

- have data points $(x_1, y_1), (x_2, y_2), (x_3, y_3)$ from a system
- need to predict y given new x
- don't really know how the system works

Your task: Make subjective guess of likely y, given $x = 0.8$



Did you guess something like this?



If so, why?

Maybe our brain picks up a pattern

- The blue dots lie roughly on a straight line
- Instictively predict red dot will lie on same line

Perhaps our brains trained/hard-wired to think in straight lines

Although thought process is subconscious, we are bringing our own assumptions into play

To guess output from previously unobserved input, must make some assumptions as to functional relationship

Assumptions may come from

- models
- subconscious preference for straight lines
- etc.

Summary:

- Statistical techniques involve assumptions
- Good assumptions lead to successful generalization

Ideally, assumptions should be based on sound theory

Subconscious feelings about straight lines probably not as good:

> *Stocks have reached what looks like a permanently high plateau. – Irving Fisher, 1929*

Informally, the rule is

$$\text{statistical learning} = \text{prior knowledge} + \text{data}$$

Ideal case:

- Lots of prior knowledge based on sound theory
- Extra structure means data has to do less work

Common cases:

- Not much prior knowledge
- Uncertain about assumptions: prior "knowledge" faulty?

We'll see that different cases call for different techniques

Statistics

Suppose that we have data $x_1, \ldots, x_N$

Def: A **statistic** is an observable function of the data

Examples:

- **Sample mean**

$$\bar{x}_N =: \bar{x} := \frac{1}{N} \sum_{n=1}^{N} x_n$$

- **Sample variance**

$$s_N^2 =: s^2 := \frac{1}{N-1} \sum_{n=1}^{N} (x_n - \bar{x})^2$$

- **Sample standard deviation**

$$s_N :=: s := \sqrt{s^2} = \left[\frac{1}{N-1} \sum_{n=1}^{N} (x_n - \bar{x})^2 \right]^{1/2}$$

- **k-th sample moment**

$$\frac{1}{N} \sum_{n=1}^{N} x_n^k$$

Given data $x_1, \ldots, x_N$ and $y_1, \ldots, y_N$

- **Sample covariance**

$$\frac{1}{N-1}\sum_{n=1}^{N}(x_n - \bar{x})(y_n - \bar{y})$$

- **Sample correlation**

$$\frac{\sum_{n=1}^{N}(x_n - \bar{x})(y_n - \bar{y})}{\sqrt{\sum_{n=1}^{N}(x_n - \bar{x})^2 \sum_{n=1}^{N}(y_n - \bar{y})^2}}$$

Common Statistics in R

```
> x <- rnorm(10)
> mean(x)
[1] -0.2069555
> var(x)
[1] 1.269357
> sd(x)
[1] 1.126657
> median(x)
[1] 0.02691741
> y <- rnorm(10)
> cov(x, y)
[1] 0.001906421
> cor(x, y)
[1] 0.004054976
```

Every statistic is a random variable!

Example: The sample mean is defined as

$$\bar{x} := \frac{1}{N} \sum_{n=1}^{N} x_n$$

More formally, it is

$$\bar{x}(\omega) := \frac{1}{N} \sum_{n=1}^{N} x_n(\omega) \qquad (\omega \in \Omega)$$

Thus, $\bar{x} \colon \Omega \to \mathbb{R}$, and hence $\bar{x}$ is a RV

Like all RVs, statistics have expectations, variance, etc.

Example: Suppose $\{x_n\}_{n=1}^{\infty} \overset{\text{IID}}{\sim} F$

Consider sample mean $\bar{x}$

From linearity of expectations,

$$\mathbb{E}[\bar{x}] = \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N} x_n\right] = \frac{1}{N}\sum_{n=1}^{N} \mathbb{E}[x_n] = \int sF(ds)$$

Even if F unknown, this tells us that $\bar{x}$ is "unbiased" for mean

Reminders:

- Please get fresh copy of course notes
- First assignment to be posted next week