2. Getting Started with Statistics

Dave Goldsman

H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

3/2/20

Outline

Georgia Tech

Lesson 2.1 — Introduction to Descriptive Statistics

What's Coming Up:

- Three high-level lessons on what Statistics is (not involving much math).
- Several lessons on estimating parameters of probability distributions.
- One lesson on certain distributions that will come up in subsequent Statistics modules — normal, time for t, χ^2, and F.

Statistics forms a rational basis for decision-making using observed or experimental **data**. We make these decisions in the face of uncertainty.

Statistics helps us answer questions concerning:

- The analysis of one population (or system).
- The comparison of many populations.

Georgia Tech

Examples:

- Election polling.
- Coke vs. Pepsi.
- The effect of cigarette smoking on the probability of getting cancer.
- The effect of a new drug on the probability of contracting hepatitis.
- What's the most popular TV show during a certain time period?
- The effect of various heat-treating methods on steel tensile strength.
- Which fertilizers improve crop yield?
- King of Siam — etc., etc., etc.

Idea (Election polling example): We can't poll every single voter. Thus, we take a **sample** of data from the **population** of voters, and try to make a reasonable conclusion based on that sample.

Statistics tells us how to conduct the sampling (i.e., how many observations to take, how to take them, etc.), and then how to draw conclusions from the sampled data.

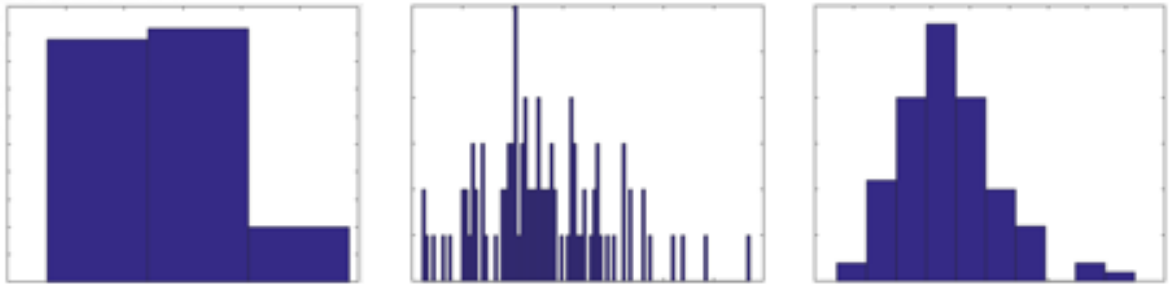
Types of Data

- **Continuous variables:** Can take on any real value in a certain interval. For example, the lifetime of a lightbulb or the weight of a newborn child.
- **Discrete variables:** Can only take on specific values. E.g., the number of accidents this week at a factory or the possible rolls of a pair of dice.
- **Categorical variables:** These data are not typically numerical. What's your favorite TV show during a certain time slot?

Plotting Data

A picture is worth 1000 words. Always plot data before doing anything else, if only to identify any obvious issues such as nonstandard distributions, missing data points, outliers, etc.

Histograms provide a quick, succinct look at what you are dealing with. If you take enough observations, the histogram will eventually converge to the true distribution. But sometimes choosing the optimal number of cells is a little tricky — like Goldilocks!



Outline

Georgia Tech

Lesson 2.2 — Summarizing Data

In addition to plotting data, how do we **summarize data?**

It's nice to have lots of data. But sometimes it's too much of a good thing! Need to summarize.

Example: Grades on a test (i.e., raw data):

23	62	91	83	82	64	73	94	94	52
67	11	87	99	37	62	40	33	80	83
99	90	18	73	68	75	75	90	36	55

Georgia Tech

Stem-and-Leaf Diagram of grades. Easy way to write down all of the data. Saves some space, and looks like a sideways histogram.

9	9944100
8	73320
7	5533
6	87422
5	52
4	0
3	763
2	3
1	81

Grouped Data

Range	Freq.	Cumul. Freq.	Proportion of observations so far
0–20	2	2	2/30
21–40	5	7	7/30
41–60	2	9	9/30
61–80	10	19	19/30
81–100	11	30	1

Summary Statistics:

$n = 30$ observations.

If X_i is the ith score, then the **sample mean** is

$$\bar{X} \equiv \sum_{i=1}^{n} X_i/n = 66.5.$$

The **sample variance** is

$$S^2 \equiv \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = 630.6.$$

Remark: Before you take any observations, $\bar{X}$ and S^2 must be regarded as *random variables*.

In general, suppose that we sample iid data $X_1, \ldots, X_n$ from the population of interest.

Example: X_i is the lifespan of the ith lightbulb we observe.

We're most interested in measuring the "center" and "spread" of the underlying distribution of the data.

Measures of Central Tendency:

Sample Mean: $\bar{X} = \sum_{i=1}^{n} X_i/n$.

Sample Median: The "middle" observation when the X_i's are arranged numerically.

Example: 16, 7, 83 gives a median of 16.

Example: 16, 7, 83, 20 gives a "reasonable" median of $\frac{16+20}{2} = 18$.

Remark: The sample median is less susceptible to "outlier" data than the sample mean. One bad number can spoil the sample mean's entire day.

Example: 7, 7, 7, 672, 7 results in a sample mean of 140 and a sample median of 7.

Sample Mode: "Most common" value. Not the most useful measure sometimes.

Example: 16, 7, 20, 83, 7 gives a mode of 7.

Measures of Variation (dispersion, spread)

Sample Variance:

$$S^2 \equiv \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right),$$

the latter expression being easier to compute.

Sample Standard Deviation: $S = +\sqrt{S^2}$.

Sample Range: $\max_i X_i - \min_i X_i$.

Remark: Suppose the data takes p different values $X_1, \ldots, X_p$, with frequencies $f_1, \ldots, f_p$, respectively.

How to calculate $\bar{X}$ and S^2 quickly?

$$\bar{X} = \sum_{j=1}^{p} f_j X_j / n \quad \text{and} \quad S^2 = \frac{\sum_{j=1}^{p} f_j X_j^2 - n\bar{X}^2}{n-1}.$$

Example: Suppose we roll a die 10 times.

X_j	1	2	3	4	5	6
f_j	2	1	1	3	0	3

Then $\bar{X} = (2 \cdot 1 + 1 \cdot 2 + \cdots + 3 \cdot 6)/10 = 3.7$, and $S^2 = 3.789$. □

Remark: If the individual observations can't be determined in frequency distributions, you might just break the observations up into c intervals.

Example: Suppose $c = 3$, where we denote the midpoint of the jth interval by m_j, $j = 1, \ldots, c$, and the total sample size $n = \sum_{j=1}^{c} f_j = 30$.

X_j interval	m_j	f_j
100–150	125	10
150–200	175	15
200–300	250	5

$$\bar{X} \approx \frac{\sum_{j=1}^{c} f_j m_j}{n} = 170.833 \quad \text{and}$$

$$S^2 \approx \frac{\sum_{j=1}^{c} f_j m_j^2 - n\bar{X}^2}{n-1} = 1814. \quad \square$$

Outline

Georgia Tech

Lesson 2.3 — Candidate Distributions

Time to make an informed guess about the type of probability distribution we're dealing with. We'll look at more-formal methodology for fitting distributions later in the course when we do goodness-of-fit tests. But for now, some preliminary things we should think about:

- Is the data from a discrete, continuous, or mixed distribution?
- Univariate/multivariate?
- How much data is available?
- Are experts around to ask about nature of the data?
- What if we do not have much/any data — can we at least guess at a good distribution?

If the distribution is a **discrete** random variable, then we have a number of familiar choices to select from.

- Bernoulli(p) (success with probability p)
- Binomial(n, p) (number of successes in n Bern(p) trials)
- Geometric(p) (number of Bern(p) trials until first success)
- Negative Binomial (number of Bern(p) trials until multiple successes)
- Poisson(λ) (counts the number of arrivals over time)
- Empirical (the all-purpose "sample" distribution based on the histogram)

If the data suggest a **continuous** distribution. . . .

- Uniform (not much is known from the data, except perhaps the minimum and maximum possible values)
- Triangular (at least we have an idea regarding the minimum, maximum, and "most likely" values)
- Exponential(λ) (e.g., interarrival times from a Poisson process)
- Normal (a good model for heights, weights, IQs, sample means, etc.)
- Beta (good for specifying bounded data)
- Gamma, Weibull, Gumbel, lognormal (reliability data)
- Empirical (our all-purpose friend)

Outline

Lesson 2.4 — Introduction to Estimation

Definition: A **statistic** is a function of the observations $X_1, \ldots, X_n$, and not explicitly dependent on any unknown parameters.

Examples of statistics: $\bar{X}$ and S^2, but not $(\bar{X} - \mu)/\sigma$.

Statistics are *random variables*. If we take two different samples, we'd expect to get two different values of a statistic.

A statistic is usually used to estimate some unknown **parameter** from the underlying probability distribution of the X_i's.

Examples of parameters: μ, σ^2.

Georgia Tech

Let $X_1, \ldots, X_n$ be iid RV's and let $T(\mathbf{X}) \equiv T(X_1, \ldots, X_n)$ be a statistic based on the X_i's. Suppose we use $T(\mathbf{X})$ to estimate some unknown parameter θ. Then $T(\mathbf{X})$ is called a **point estimator** for θ.

Examples: $\bar{X}$ is usually a point estimator for the mean $\mu = \mathrm{E}[X_i]$, and S^2 is often a point estimator for the variance $\sigma^2 = \mathrm{Var}(X_i)$.

It would be nice if $T(\mathbf{X})$ had certain properties:

- Its expected value should equal the parameter it's trying to estimate.
- It should have low variance.

Outline

Georgia Tech

Lesson 2.5 — Unbiased Estimation

Definition: $T(\mathbf{X})$ is **unbiased** for θ if $\mathrm{E}[T(\mathbf{X})] = \theta$.

Example/Theorem: Suppose $X_1, \ldots, X_n$ are iid anything with mean μ. Then $\bar{X}$ is always unbiased for μ.

$$\mathrm{E}[\bar{X}] \;=\; \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \;=\; \mathrm{E}[X_i] \;=\; \mu.$$

That's why $\bar{X}$ is called the **sample mean**. □

Baby Example: In particular, suppose $X_1, \ldots, X_n$ are iid Exp(λ). Then $\bar{X}$ is unbiased for $\mu = \mathrm{E}[X_i] = 1/\lambda$.

But be careful.... $1/\bar{X}$ is *biased* for λ in this exponential case, i.e., $\mathrm{E}[1/\bar{X}] \neq 1/\mathrm{E}[\bar{X}] = \lambda$. □

Georgia Tech

Example/Theorem: Suppose $X_1, \ldots, X_n$ are iid anything with mean μ and variance σ^2. Then S^2 is always unbiased for σ^2.

$$\mathrm{E}[S^2] \;=\; \mathrm{E}\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\right] \;=\; \mathrm{Var}(X_i) \;=\; \sigma^2.$$

This is why S^2 is called the **sample variance**. $\square$

Baby Example: Suppose $X_1, \ldots, X_n$ are iid Exp(λ). Then S^2 is unbiased for $\mathrm{Var}(X_i) = 1/\lambda^2$. $\square$

Proof (of general result): First, some algebra gives

$$\begin{aligned}
\sum_{i=1}^{n}(X_i - \bar{X})^2 &= \sum_{i=1}^{n}(X_i^2 - 2\bar{X}X_i + \bar{X}^2) \\
&= \sum_{i=1}^{n} X_i^2 - 2\bar{X}\sum_{i=1}^{n} X_i + n\bar{X}^2 \\
&= \sum_{i=1}^{n} X_i^2 - 2n\bar{X}^2 + n\bar{X}^2 \\
&= \sum_{i=1}^{n} X_i^2 - n\bar{X}^2.
\end{aligned}$$

So...

$$
\begin{aligned}
\mathrm{E}[S^2] &= \frac{1}{n-1}\mathrm{E}\Big[\sum_{i=1}^{n}(X_i-\bar{X})^2\Big] = \frac{1}{n-1}\mathrm{E}\Big[\sum_{i=1}^{n}X_i^2 - n\bar{X}^2\Big] \\
&= \frac{1}{n-1}\Big(\sum_{i=1}^{n}\mathrm{E}[X_i^2] - n\mathrm{E}[\bar{X}^2]\Big) \\
&= \frac{n}{n-1}\Big(\mathrm{E}[X_1^2] - \mathrm{E}[\bar{X}^2]\Big) \quad \text{(since the } X_i\text{'s are iid)} \\
&= \frac{n}{n-1}\Big(\mathrm{Var}(X_1) + (\mathrm{E}[X_1])^2 - \mathrm{Var}(\bar{X}) - (\mathrm{E}[\bar{X}])^2\Big) \\
&= \frac{n}{n-1}(\sigma^2 - \sigma^2/n) \quad \text{(since } \mathrm{E}[X_1] = \mathrm{E}[\bar{X}] \text{ and } \mathrm{Var}(\bar{X}) = \sigma^2/n\text{)} \\
&= \sigma^2. \quad \text{Done.} \quad \square
\end{aligned}
$$

Remark: S is *not* unbiased for the standard deviation σ.

Big Example: Suppose that $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Unif}(0, \theta)$, i.e., the pdf is $f(x) = 1/\theta$, for $0 < x < \theta$. Think of it this way: I give you a bunch of random numbers between 0 and θ, and you have to guess what θ is.

We'll look at *three* unbiased estimators for θ:

$Y_1 = 2\bar{X}$.

$Y_2 = \frac{n+1}{n} \max_{1 \le i \le n} X_i$.

$$Y_3 = \begin{cases} 12\bar{X} & \text{w.p. } 1/2 \\ -8\bar{X} & \text{w.p. } 1/2. \end{cases}$$

If they're all unbiased, which one's the best?

"Good" Estimator: $Y_1 = 2\bar{X}$.

Proof (that it's unbiased): $\mathrm{E}[Y_1] = 2\mathrm{E}[\bar{X}] = 2\mathrm{E}[X_i] = \theta$. $\square$

"Better" Estimator: $Y_2 = \frac{n+1}{n} \max_{1 \le i \le n} X_i$.

Why might this estimator for θ make sense? (We'll say why it's "better" in a little while.)

Proof (that it's unbiased): $\mathrm{E}[Y_2] = \frac{n+1}{n}\mathrm{E}[\max_i X_i] = \theta$ iff

$$\mathrm{E}[\max X_i] = \frac{n\theta}{n+1} \quad \text{(which is what we'll show below).}$$

First, let's get the cdf of $M \equiv \max_i X_i$:

$$
\begin{aligned}
P(M \le y) &= P(X_1 \le y \text{ and } X_2 \le y \text{ and } \cdots \text{ and } X_n \le y) \\
&= P(X_1 \le y)P(X_2 \le y)\cdots P(X_n \le y) \quad (X_i\text{'s indep}) \\
&= [P(X_1 \le y)]^n \quad (X_i\text{'s indentically distributed}) \\
&= \left[\int_0^y f_{X_1}(x)\,dx\right]^n \\
&= \left[\int_0^y (1/\theta)\,dx\right]^n \\
&= (y/\theta)^n.
\end{aligned}
$$

This implies that the pdf of M is

$$f_M(y) \equiv \frac{d}{dy}(y/\theta)^n = \frac{ny^{n-1}}{\theta^n}, \quad 0 < y < \theta,$$

and this implies that

$$\mathrm{E}[M] = \int_0^\theta y f_M(y)\, dy = \int_0^\theta \frac{ny^n}{\theta^n}\, dy = \frac{n\theta}{n+1}.$$

Whew! This finally shows that $Y_2 = \frac{n+1}{n} \max_{1 \le i \le n} X_i$ is an unbiased estimator for θ! $\square$

Lastly, let's look at...

"Ugly" Estimator:

$$Y_3 = \begin{cases} 12\bar{X} & \text{w.p. } 1/2 \\ -8\bar{X} & \text{w.p. } 1/2. \end{cases}$$

Ha! It's possible to get a *negative* estimate for θ, which is strange since $\theta > 0$!

Proof (that it's unbiased):

$$\mathrm{E}[Y_3] = 12\mathrm{E}[\bar{X}] \cdot \frac{1}{2} - 8\mathrm{E}[\bar{X}] \cdot \frac{1}{2} = 2\mathrm{E}[\bar{X}] = \theta. \quad \square$$

Usually, it's *good* for an estimator to be unbiased, but the "ugly" estimator Y_3 shows that unbiased estimators can sometimes be goofy.

Therefore, let's look at some other properties an estimator can have.

For instance, consider the *variance* of an estimator.

Big Example (cont'd): Again suppose that

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Unif}(0, \theta).$$

Recall that both $Y_1 = 2\bar{X}$ and $Y_2 = \frac{n+1}{n} M$ are unbiased for θ.

Let's find $\text{Var}(Y_1)$ and $\text{Var}(Y_2)$. First,

$$\text{Var}(Y_1) = 4\text{Var}(\bar{X}) = \frac{4}{n} \cdot \text{Var}(X_i) = \frac{4}{n} \cdot \frac{\theta^2}{12} = \frac{\theta^2}{3n}.$$

Meanwhile,

$$\begin{aligned}
\mathrm{Var}(Y_2) &= \left(\frac{n+1}{n}\right)^2 \mathrm{Var}(M) \\
&= \left(\frac{n+1}{n}\right)^2 \mathrm{E}[M^2] - \left(\frac{n+1}{n} \cdot \mathrm{E}[M]\right)^2 \\
&= \left(\frac{n+1}{n}\right)^2 \int_0^\theta \frac{n y^{n+1}}{\theta^n}\, dy \; - \; \theta^2 \\
&= \theta^2 \cdot \frac{(n+1)^2}{n(n+2)} - \theta^2 \; = \; \frac{\theta^2}{n(n+2)} \; < \; \frac{\theta^2}{3n}.
\end{aligned}$$

Thus, both Y_1 and Y_2 are unbiased, but Y_2 has *much lower variance* than Y_1. We can break the "unbiasedness tie" by choosing Y_2. $\square$

Outline

Lesson 2.6 — Mean Squared Error

We'll now talk about a statistical performance measure that combines information about the bias and the variance of an estimator.

Definition: The **Mean Squared Error (MSE)** of an estimator $T(\mathbf{X})$ of θ is

$$\text{MSE}(T(\mathbf{X})) \equiv \text{E}[(T(\mathbf{X}) - \theta)^2].$$

Before giving an easier interpretation of MSE, define the **bias** of an estimator for the parameter θ,

$$\text{Bias}(T(\mathbf{X})) \equiv \text{E}[T(\mathbf{X})] - \theta.$$

Theorem/Proof: Easier interpretation of MSE.

$$\begin{aligned}
\text{MSE}(T(\mathbf{X})) &= \text{E}[(T(\mathbf{X}) - \theta)^2] \\
&= \text{E}[T^2] - 2\theta\text{E}[T] + \theta^2 \\
&= \text{E}[T^2] - (\text{E}[T])^2 + (\text{E}[T])^2 - 2\theta\text{E}[T] + \theta^2 \\
&= \text{Var}(T) + \underbrace{(\text{E}[T] - \theta)}_{\text{Bias}}{}^2.
\end{aligned}$$

So MSE $=$ Bias^2 $+$ Var, and thus combines the bias and variance of an estimator. $\square$

The lower the MSE the better. If $T_1(\mathbf{X})$ and $T_2(\mathbf{X})$ are two estimators of θ, we'd usually prefer the one with the lower MSE — even if it happens to have higher bias.

Definition: The **relative efficiency** of $T_2(\mathbf{X})$ to $T_1(\mathbf{X})$ is $\text{MSE}(T_1(\mathbf{X}))/\text{MSE}(T_2(\mathbf{X}))$. If this quantity is < 1, then we'd want $T_1(\mathbf{X})$.

Example: Suppose that estimator A has bias = 3 and variance = 10, while estimator B has bias $= -2$ and variance = 14. Which estimator (A or B) has the lower mean squared error?

Solution: $\text{MSE} = \text{Bias}^2 + \text{Var}$, so

$$\text{MSE(A)} = 9 + 10 = 19 \qquad \text{and} \qquad \text{MSE(B)} = 4 + 14 = 18.$$

Thus, B has lower MSE. $\square$

Example: $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Unif}(0, \theta)$.

Two estimators: $Y_1 = 2\bar{X}$, and $Y_2 = \frac{n+1}{n} \max_i X_i$.

Showed before $\text{E}[Y_1] = \text{E}[Y_2] = \theta$ (so both estimators are unbiased).

Also, $\text{Var}(Y_1) = \frac{\theta^2}{3n}$, and $\text{Var}(Y_2) = \frac{\theta^2}{n(n+2)}$.

Thus,

$$\text{MSE}(Y_1) = \frac{\theta^2}{3n} \quad \text{and} \quad \text{MSE}(Y_2) = \frac{\theta^2}{n(n+2)},$$

so Y_2 is better (by an order of magnitude, actually). $\square$

Outline

Georgia Tech

Lesson 2.7 — Maximum Likelihood Estimation

Definition: Consider an iid random sample $X_1, \ldots, X_n$, where each X_i has pmf/pdf $f(x)$. Further, suppose that θ is some unknown parameter from X_i.

The **likelihood function** is $L(\theta) \equiv \prod_{i=1}^{n} f(x_i)$.

The **maximum likelihood estimator (MLE)** of θ is the value of θ that maximizes $L(\theta)$. The MLE is a function of the X_i's and is a RV.

Remark: We can very informally regard the MLE as the "most likely" estimate of θ.

Example: Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Exp}(\lambda)$. Find the MLE for λ.

First of all, the likelihood function is

$$L(\lambda) = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n \exp\Big(-\lambda \sum_{i=1}^{n} x_i\Big).$$

Now maximize $L(\lambda)$ with respect to λ. Could take the derivative and plow through all of the horrible algebra. Too tedious. Need a trick. . . .

Useful Trick: Since the natural log function is one-to-one, it's easy to see that the λ that maximizes $L(\lambda)$ also maximizes $\ell\text{n}(L(\lambda))$!

$$\ell\text{n}(L(\lambda)) = \ell\text{n}\Big(\lambda^n \exp\Big(-\lambda \sum_{i=1}^{n} x_i\Big)\Big) = n\ell\text{n}(\lambda) - \lambda \sum_{i=1}^{n} x_i.$$

The trick makes our job less horrible.

$$\frac{d}{d\lambda}\ell\mathrm{n}(L(\lambda)) \;=\; \frac{d}{d\lambda}\left(n\ell\mathrm{n}(\lambda) - \lambda\sum_{i=1}^{n} x_i\right) \;=\; \frac{n}{\lambda} - \sum_{i=1}^{n} x_i \;\equiv\; 0.$$

This implies that the MLE is $\hat{\lambda} = 1/\bar{X}$. $\square$

Remarks:

- $\hat{\lambda} = 1/\bar{X}$ makes sense, since $\mathrm{E}[X] = 1/\lambda$.
- At the end, we put a little $\widehat{\mathrm{hat}}$ over λ to indicate that this is the MLE. It's like a party hat!
- At the end, we make all of the little x_i's into big X_i's to indicate that this is a random variable.
- Just to be careful, you "probably" ought to do a second-derivative test.

Example: Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bern}(p)$. Find the MLE for p.
Useful trick for this problem: Since

$$X_i = \begin{cases} 1 & \text{w.p. } p \\ 0 & \text{w.p. } 1-p, \end{cases}$$

we can write the pmf as

$$f(x) = p^x(1-p)^{1-x}, \quad x = 0, 1.$$

Thus, the likelihood function is

$$L(p) = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}.$$

This implies that

$$\ell\mathrm{n}(L(p)) \;=\; \sum_{i=1}^{n} x_i\,\ell\mathrm{n}(p) + \Big(n - \sum_{i=1}^{n} x_i\Big)\ell\mathrm{n}(1-p)$$

$\Rightarrow$

$$\frac{d}{dp}\ell\mathrm{n}(L(p)) \;=\; \frac{\sum_i x_i}{p} - \frac{n - \sum_i x_i}{1-p} \;\equiv\; 0$$

$\Rightarrow$

$$(1-p)\Big(\sum_{i=1}^{n} x_i\Big) \;=\; p\Big(n - \sum_{i=1}^{n} x_i\Big)$$

$\Rightarrow$

$$\hat{p} \;=\; \bar{X}.$$

This makes sense since $\mathrm{E}[X] = p$. $\square$

Outline

Georgia Tech

Lesson 2.8 — Trickier MLE Examples

Example: $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Nor}(\mu, \sigma^2)$. Get *simultaneous* MLEs for μ and σ^2.

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2} \frac{(x_i - \mu)^2}{\sigma^2} \right\}$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\sigma^2} \right\}.$$

$$\Rightarrow \quad \ell\text{n}(L(\mu, \sigma^2)) = -\frac{n}{2}\ell\text{n}(2\pi) - \frac{n}{2}\ell\text{n}(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2$$

$$\Rightarrow \quad \frac{\partial}{\partial \mu} \ell\text{n}(L(\mu, \sigma^2)) = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) \equiv 0,$$

and so $\hat{\mu} = \bar{X}$.

Similarly, take the partial with respect to σ^2 (*not* σ),

$$\frac{\partial}{\partial \sigma^2} \ell \mathrm{n}(L(\mu, \sigma^2)) \;=\; -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (x_i - \hat{\mu})^2 \;\equiv\; 0,$$

and eventually get

$$\widehat{\sigma^2} \;=\; \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2. \quad \square$$

Remark: Notice how close $\widehat{\sigma^2}$ is to the (unbiased) sample variance,

$$S^2 \;=\; \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 \;=\; \frac{n}{n-1} \widehat{\sigma^2}.$$

$\widehat{\sigma^2}$ is a little bit biased, but it has slightly less variance than S^2. Anyway, as n gets big, S^2 and $\widehat{\sigma^2}$ become the same.

Example: The pdf of the Gamma distribution w/parameters r and λ is

$$f(x) \;=\; \frac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-\lambda x}, \quad x > 0.$$

Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Gam}(r, \lambda)$. Find the MLEs for r and λ.

$$L(r,\lambda) \;=\; \prod_{i=1}^n f(x_i) \;=\; \frac{\lambda^{nr}}{[\Gamma(r)]^n} \Big(\prod_{i=1}^n x_i \Big)^{r-1} e^{-\lambda \sum_i x_i}$$

$$\Rightarrow \quad \ell\text{n}(L) \;=\; rn\,\ell\text{n}(\lambda) - n\,\ell\text{n}(\Gamma(r)) + (r-1)\ell\text{n}\Big(\prod_{i=1}^n x_i \Big) - \lambda \sum_{i=1}^n x_i$$

$$\Rightarrow \quad \frac{\partial}{\partial \lambda} \ell\text{n}(L) \;=\; \frac{rn}{\lambda} - \sum_{i=1}^n x_i \;\equiv\; 0,$$

so that $\hat{\lambda} = \hat{r}/\bar{X}$.

The Trouble in River City is, we need to find $\hat{r}$. To do so, we have

$$\begin{aligned}
\frac{\partial}{\partial r}\ell\mathrm{n}(L) &= \frac{\partial}{\partial r}\left[rn\,\ell\mathrm{n}(\lambda) - n\,\ell\mathrm{n}(\Gamma(r)) + (r-1)\ell\mathrm{n}\Big(\prod_{i=1}^{n} x_i\Big) - \lambda\sum_{i=1}^{n} x_i\right] \\
&= n\,\ell\mathrm{n}(\lambda) - \frac{n}{\Gamma(r)}\,\frac{d}{dr}\Gamma(r) + \ell\mathrm{n}\Big(\prod_{i=1}^{n} x_i\Big) \\
&= n\,\ell\mathrm{n}(\lambda) - n\Psi(r) + \ell\mathrm{n}\Big(\prod_{i=1}^{n} x_i\Big) \;\equiv\; 0,
\end{aligned}$$

where $\Psi(r) \equiv \Gamma'(r)/\Gamma(r)$ is the **digamma function.**

At this point, substitute in $\hat{\lambda} = \hat{r}/\bar{X}$, and use a *computer method* (bisection, Newton's method, etc.) to search for the value of r that solves

$$n\,\ell\mathrm{n}(r/\bar{X}) - n\Psi(r) + \ell\mathrm{n}\Big(\prod_{i=1}^{n} x_i\Big) \;\equiv\; 0.$$

The gamma function is readily available in any reasonable software package; but if the digamma function happens to be unavailable in your town, you can take advantage of the approximation

$$\Gamma'(r) \;\doteq\; \frac{\Gamma(r+h) - \Gamma(r)}{h} \quad \text{(for any small } h \text{ of your choosing).} \quad \square$$

Example: Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Unif}(0, \theta)$. Find the MLE for θ.

The pdf is $f(x) = 1/\theta$, $0 < x < \theta$, (beware of the funny limits). Then

$$L(\theta) \;=\; \prod_{i=1}^{n} f(x_i) \;=\; 1/\theta^n \quad \text{if } 0 \le x_i \le \theta, \forall i$$

In order to have $L(\theta) > 0$, we must have $0 \le x_i \le \theta$, $\forall i$. In other words, we must have $\theta \ge \max_i x_i$.

Subject to this constraint, $L(\theta) = 1/\theta^n$ is maximized at the smallest possible θ value, namely, $\hat{\theta} = \max_i X_i$.

This makes sense in light of the similar (unbiased) estimator, $Y_2 = \frac{n+1}{n} \max_i X_i$, from a previous lesson. $\square$

Remark: We used very little calculus in this example!

Outline

Georgia Tech

Lesson 2.9 — Invariance Property of MLEs

We can get MLEs of functions of parameters almost for free!

Theorem (Invariance Property): If $\hat{\theta}$ is the MLE of some parameter θ and $h(\cdot)$ is any reasonable function, then $h(\hat{\theta})$ is the MLE of $h(\theta)$.

Remark: We noted before that such a property does *not* hold for unbiasedness. For instance, although $\mathrm{E}[S^2] = \sigma^2$, it is usually the case that $\mathrm{E}[\sqrt{S^2}] \neq \sigma$.

Remark: The proof of the Invariance Property is "easy" when $h(\cdot)$ is a one-to-one function. It's not so easy — but still generally true — when $h(\cdot)$ is nastier.

Example: Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Nor}(\mu, \sigma^2)$.

We saw that the MLE for σ^2 is $\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$.

If we consider the function $h(y) = +\sqrt{y}$, then the Invariance Property says that the MLE of σ is

$$\hat{\sigma} = \sqrt{\widehat{\sigma^2}} = \sqrt{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2}. \quad \square$$

Example: Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bern}(p)$.

We saw that the MLE for p is $\hat{p} = \bar{X}$. Then Invariance says that the MLE for $\text{Var}(X_i) = p(1-p)$ is $\hat{p}(1-\hat{p}) = \bar{X}(1-\bar{X})$. $\square$

Example: Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Exp}(\lambda)$.

We define the **survival function** as

$$\bar{F}(x) \;=\; P(X > x) \;=\; 1 - F(x) \;=\; e^{-\lambda x}.$$

In addition, we saw that the MLE for λ is $\hat{\lambda} = 1/\bar{X}$.

Then Invariance says that the MLE of $\bar{F}(x)$ is

$$\widehat{\bar{F}(x)} \;=\; e^{-\hat{\lambda} x} \;=\; e^{-x/\bar{X}}.$$

This kind of thing is used all of the time in the actuarial sciences. □

Outline

Georgia Tech

Lesson 2.10 — Method of Moments Estimation

Recall that the kth **moment** of a random variable X is

$$\mu_k \equiv \mathrm{E}[X^k] = \begin{cases} \sum_x x^k f(x) & \text{if } X \text{ is discrete} \\ \int_{\mathbb{R}} x^k f(x)\, dx & \text{if } X \text{ is continuous.} \end{cases}$$

Definition: Suppose $X_1, \ldots, X_n$ are iid random variables. Then the **method of moments (MoM) estimator** for μ_k is $m_k \equiv \sum_{i=1}^n X_i^k/n$.

Remark: As $n \to \infty$, the Law of Large Numbers implies that $\sum_{i=1}^n X_i^k/n \to \mathrm{E}[X^k]$, i.e., $m_k \to \mu_k$ (so this is a good estimator).

Remark: You should always love your MoM!

Examples:

The MoM estimator for the true mean $\mu_1 = \mu = \mathrm{E}[X_i]$ is the sample mean $m_1 = \bar{X} = \sum_{i=1}^n X_i/n$.

The MoM estimator for $\mu_2 = \mathrm{E}[X_i^2]$ is $m_2 = \sum_{i=1}^n X_i^2/n$.

The MoM estimator for $\mathrm{Var}(X_i) = \mathrm{E}[X_i^2] - (\mathrm{E}[X_i])^2 = \mu_2 - \mu_1^2$ is

$$m_2 - m_1^2 \;=\; \frac{1}{n}\sum_{i=1}^n X_i^2 - \bar{X}^2 \;=\; \frac{n-1}{n}S^2.$$

(For large n, it's also OK to use S^2.)

General Game Plan: Express the parameter of interest in terms of the true moments $\mu_k = \mathrm{E}[X^k]$. Then substitute in the sample moments m_k.

Example: Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Pois}(\lambda)$.

Since $\lambda = \text{E}[X_i]$, a MoM estimator for λ is $\bar{X}$.

But also note that $\lambda = \text{Var}(X_i)$, so another MoM estimator for λ is $\frac{n-1}{n}S^2$ (or plain old S^2). $\square$

Usually use the easier-looking estimator if you have a choice.

Example: Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Nor}(\mu, \sigma^2)$.

MoM estimators for μ and σ^2 are $\bar{X}$ and $\frac{n-1}{n}S^2$ (or S^2), respectively.

For this example, these estimators are the same as the MLEs. $\square$

Let's finish up with a less-trivial example....

Example: Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Beta}(a, b)$. The pdf is

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}, \quad 0 < x < 1.$$

It turns out (after lots of algebra) that

$$\text{E}[X] = \frac{a}{a+b} \quad \text{and} \quad \text{Var}(X) = \frac{ab}{(a+b)^2(a+b+1)}.$$

Let's estimate a and b via MoM.

We have

$$\mathrm{E}[X] \;=\; \frac{a}{a+b} \;\Rightarrow\; a \;=\; \frac{b\,\mathrm{E}[X]}{1-\mathrm{E}[X]} \;\doteq\; \frac{b\bar{X}}{1-\bar{X}}, \tag{1}$$

so

$$\mathrm{Var}(X) \;=\; \frac{ab}{(a+b)^2(a+b+1)} \;=\; \frac{\mathrm{E}[X]b}{(a+b)(a+b+1)}.$$

Plug into the above $\bar{X}$ for $\mathrm{E}[X]$, S^2 for $\mathrm{Var}(X)$, and $\frac{b\bar{X}}{1-\bar{X}}$ for a. Then after lots of algebra, we can solve for b:

$$b \;\doteq\; \frac{(1-\bar{X})^2\bar{X}}{S^2} - 1 + \bar{X}.$$

To finish up, you can plug back into Equation (1) to get the MoM estimator for a.

Example: Consider the following data set consisting of $n = 10$ observations that we have obtained from a Beta distribution.

0.86 0.77 0.84 0.38 0.83 0.54 0.77 0.94 0.37 0.40

We immediately have $\bar{X} = 0.67$, and $S^2 = 0.04971$. Then the MoM estimators are

$$b \doteq \frac{(1-\bar{X})^2\bar{X}}{S^2} - 1 + \bar{X} = 1.1377,$$

and then

$$a \doteq \frac{b\bar{X}}{1-\bar{X}} = 2.310. \quad \square$$

Outline

Georgia Tech

Introduction and Normal Distribution

Goal: Talk about some distributions we'll need later to do "confidence intervals" (CIs) and "hypothesis tests": Normal, χ^2, t, and F.

Definition: Recall that a **statistic** is just a function of the observations $X_1, \ldots, X_n$ from a random sample. The function does not depend explicitly on any unknown parameters.

Example: $\bar{X}$ and S^2 are statistics, but $(\bar{X} - \mu)/\sigma$ is not.

Since statistics are RV's, it's useful to figure out their distributions.
The distribution of a statistic is called a **sampling distribution**.

Example: $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Nor}(\mu, \sigma^2) \Rightarrow \bar{X} \sim \text{Nor}(\mu, \sigma^2/n)$.

The normal is used to get CIs and do hypothesis tests for μ.

Georgia Tech

χ^2 Distribution

Definition/Theorem: If $Z_1, \ldots, Z_k \overset{\text{iid}}{\sim} \text{Nor}(0,1)$, then $Y \equiv \sum_{i=1}^{k} Z_i^2$ has the **chi-squared distribution** with ***k* degrees of freedom (df)**, and we write $Y \sim \chi^2(k)$.

The term "df" informally corresponds to the number of "independent pieces of information" you have. For example, if you have RV's $X_1, \ldots, X_n$ such that $\sum_{i=1}^{n} X_i = c$, a known constant, then you might have $n-1$ df, since knowledge of any $n-1$ of the X_i's gives you the remaining X_i.

We also informally "lose" a degree of freedom every time we have to estimate a parameter. For instance, if we have access to n observations, but have to estimate two parameters μ and σ^2, then we might only end up with $n-2$ df.

In reality, df corresponds to the number of dimensions of a certain space (not covered in this course)!

Georgia Tech

The pdf of the chi-squared distribution is

$$f_Y(y) = \frac{1}{2^{k/2}\Gamma\left(\frac{k}{2}\right)} y^{\frac{k}{2}-1} e^{-y/2}, \quad y > 0.$$

Fun Facts: Can show that $\mathrm{E}[Y] = k$, and $\mathrm{Var}(Y) = 2k$.

The exponential distribution is a special case of the chi-squared distribution. In fact, $\chi^2(2) \sim \mathrm{Exp}(1/2)$.

Proof: Just plug $k = 2$ into the pdf. $\square$

For $k > 2$, the $\chi^2(k)$ pdf is skewed to the right. (You get an occasional "large" observation.)

For large k, the $\chi^2(k)$ is approximately normal (by the CLT).

Definition: The $(1-\alpha)$ **quantile** of a RV X is that value x_α such that $P(X > x_\alpha) = 1 - F(x_\alpha) = \alpha$. Note that $x_\alpha = F^{-1}(1-\alpha)$, where $F^{-1}(\cdot)$ is the **inverse cdf** of X.

Notation: If $Y \sim \chi^2(k)$, then we denote the $(1-\alpha)$ quantile with the special symbol $\chi^2_{\alpha,k}$ (instead of x_α). In other words, $P(Y > \chi^2_{\alpha,k}) = \alpha$. You can look up $\chi^2_{\alpha,k}$, e.g., in a table at the back of the book or via the Excel function `CHISQ.INV`$(1-\alpha, k)$.

Example: If $Y \sim \chi^2(10)$, then

$$P(Y > \chi^2_{0.05,10}) \;=\; 0.05,$$

where we can look up $\chi^2_{0.05,10} = 18.31$. $\square$

Theorem: χ^2's add up. If $Y_1, \ldots, Y_n$ are *independent* with $Y_i \sim \chi^2(d_i)$, for all i, then $\sum_{i=1}^n Y_i \sim \chi^2(\sum_{i=1}^n d_i)$.

Proof: Just use mgf's. Won't go thru it here. $\square$

So where does the χ^2 distribution come up in statistics?

It usually arises when we try to estimate σ^2.

Example: If $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Nor}(\mu, \sigma^2)$, then, as we'll show in the next module,

$$S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2 \sim \frac{\sigma^2 \chi^2(n-1)}{n-1}. \quad \square$$

t Distribution

Definition/Theorem: Suppose that $Z \sim \text{Nor}(0,1)$, $Y \sim \chi^2(k)$, and Z and Y are independent. Then $T \equiv Z/\sqrt{Y/k}$ has the **Student t distribution** with k degrees of freedom, and we write $T \sim t(k)$.

The pdf is

$$f_T(x) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{\pi k}\,\Gamma\left(\frac{k}{2}\right)}\left(\frac{x^2}{k}+1\right)^{-\frac{k+1}{2}}, \quad x \in \mathbb{R}.$$

Fun Facts: The $t(k)$ looks like the Nor(0,1), except the t has fatter tails.

The $k = 1$ case gives the **Cauchy** distribution, which has *really* fat tails.

As the degrees of freedom k becomes large, $t(k) \to \text{Nor}(0,1)$.

Can show that $\text{E}[T] = 0$ for $k > 1$, and $\text{Var}(T) = \frac{k}{k-2}$ for $k > 2$.

Georgia Tech

Notation: If $T \sim t(k)$, then we denote the $(1-\alpha)$ quantile by $t_{\alpha,k}$. In other words, $P(T > t_{\alpha,k}) = \alpha$.

Example: If $T \sim t(10)$, then $P(T > t_{0.05,10}) = 0.05$, where we find $t_{0.05,10} = 1.812$ in the back of the book or via the Excel function `T.INV(`$1-\alpha, k$`)`. $\square$

Remarks: So what do we use the t distribution for in statistics?

It's used when we find confidence intervals and conduct hypothesis tests for the mean μ. Stay tuned.

By the way, why did I originally call it the **Student** t distribution?

"Student" is the pseudonym of the guy (William Gossett) who first derived it. Gossett was a statistician at the Guinness Brewery.

F Distribution

Definition/Theorem: Suppose that $X \sim \chi^2(n)$, $Y \sim \chi^2(m)$, and X and Y are independent. Then $F \equiv \frac{X/n}{Y/m} = mX/(nY)$ has the ***F* distribution** with n and m df, denoted $F \sim F(n, m)$.

The pdf is

$$f_F(x) = \frac{\Gamma\left(\frac{n+m}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{m}{2}\right)} \frac{\left(\frac{n}{m}\right)^{\frac{n}{2}} x^{\frac{n}{2}-1}}{\left(\frac{n}{m}x + 1\right)^{\frac{n+m}{2}}}, \quad x > 0.$$

Fun Facts: The $F(n, m)$ is usually a bit skewed to the right.

Note that you have to specify two df's.

Can show that $\mathrm{E}[F] = \frac{m}{m-2}$ $(m > 2)$, and $\mathrm{Var}(F) =$ blech.

t distribution is a special case — can you figure out which?

Notation: If $F \sim F(n, m)$, then we denote the $(1 - \alpha)$ quantile by $F_{\alpha,n,m}$. That is, $P(F > F_{\alpha,n,m}) = \alpha$.

Tables came be found in back of the book for various α, n, m or you can use the Excel function `F.INV(`$1 - \alpha, n, m$`)`

Example: If $F \sim F(5, 10)$, then $P(F > F_{0.05,5,10}) = 0.05$, where we find $F_{0.05,5,10} = 3.326$. $\square$

Remarks: It can be shown that $F_{1-\alpha,m,n} = 1/F_{\alpha,n,m}$. Use this fact if you have to find something like $F_{0.95,10,5} = 1/F_{0.05,5,10} = 1/3.326$.

So what do we use the F distribution for in statistics?

It's used when we find confidence intervals and conduct hypothesis tests for the ratio of variances from two different processes. Details later.

Georgia Tech