Brooklyn College, CUNY

Math 4501 – Statistics

Lecture Notes

Spring 2018

Christian Beneš

cbenes@brooklyn.cuny.edu

DISCLAIMER: These are not meant to be a complete set of notes! They are meant to supplement to the material I will be covering in class, in particular when graphs or numerical tables are involved.

Math 4501 (Spring 2018) January 29, 2018
Prof. Christian Beneš

Lecture #1: Introduction; Probability Review

1.1 The Goals of Statistics

What are the goals of statistics?

To look at data sets and make inferences about the underlying randomness and use these inferences to make predictions.

Some types of questions that statistics can answer:

- Is the stock market ”completely random”? if not, how can we profit from it.
- Does smoking cause cancer? A way to formulate this question precisely is the following:

  “How likely is it that differences between the cancer rates of smokers and non-smokers are caused purely by chance?”

  For instance, if we are given a sample of 3 smokers and 3 non-smokers and the only person with cancer is a non-smoker, what conclusion should we draw?
- Does pumping all that CO_2 into the atmosphere cause global warming?

The problem with statistics: You never get an exact answer. The best you can say is, e.g., “It’s very likely that smoking causes cancer”.

The great thing about statistics: It’s **very** useful. In biology, geology, genomics, finance, and any other field where understanding data is important. It’s also often elegant.

Understanding statistics (not just learning a bunch of recipes) is crucial to avoid a major mis-interpretation of real-world data.

Example 1.1. (Number of games with given number of goals at the 2014 FIFA World Cup)

a	0	1	2	3	4	5	6	7	8	> 8
#(games with a goals)	7	12	8	20	9	4	2	1	1	0

This table yields the related table:

a	0	1	2	3	4	5	6	7	8	> 8
%(games with a goals)	.109375	.1875	.125	.3125	.140625	.0625	.03125	.015625	.015625	0

If one were to look for a distribution which describes these data, one would definitely hope that the mean of that distribution be the same as (or at least close to) the sample mean (to be defined formally below).

The sample mean, or in this context, the average number of goals scored per game at the 2014 Soccer World Cup, is

$$\frac{1}{64}(0 \cdot 7 + 1 \cdot 12 + 2 \cdot 8 + 3 \cdot 20 + 4 \cdot 9 + 5 \cdot 4 + 6 \cdot 2 + 7 \cdot 1 + 8 \cdot 1 = 171/64 = 2.671875,$$

where 64 is the total number of games at the 2014 FIFA World Cup.

Another way to think about this sample mean is as a *weighted average*:

$$\frac{0}{64} \cdot 7 + \frac{1}{64} \cdot 12 + \frac{2}{64} \cdot 8 + \frac{3}{64} \cdot 20 + \frac{4}{64} \cdot 9 + \frac{5}{64} \cdot 4 + \frac{6}{64} \cdot 2 + \frac{7}{64} \cdot 1 + \frac{8}{64} \cdot 1 = 171/64 = 2.671875.$$

1.2 Some Important Distributions

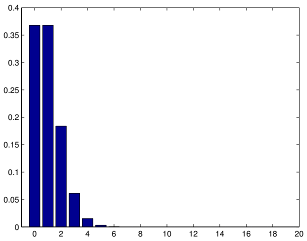
1.2.1 The Poisson Distribution

The Poisson random variable $Po(\lambda), \lambda > 0$

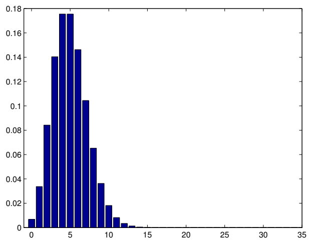
- Is used as a model in different contexts. For instance, the following usually obey a Poisson law:

    1. The number of typos on the page of a book.
    2. The number of cars passing through Times Square between 12:00 and 12:01 p.m.
    3. The number of people in U.S. jails on a given day.
    4. The number of students in this class who will ace a given exam.

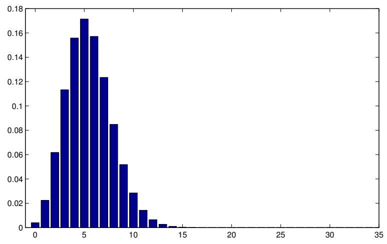
- Has p.m.f. defined by $P(X = i) = e^{-\lambda} \frac{\lambda^i}{i!}, i = 0, 1, 2, ...$
- Its mode occurs at the largest integer $k \leq \lambda$.
- Can be used to approximate the binomial if n is large and p is small (so that np is "moderate"). In that case, $Bin(n, p) \approx Po(np)$.
- Has mean λ and variance λ.



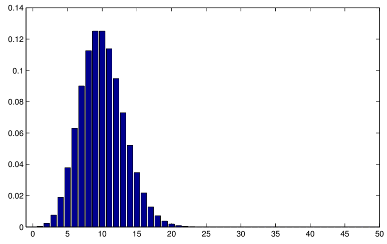
$Po(1)$



$Po(5)$



$Po(5.5)$

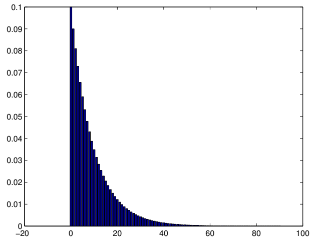


$Po(10)$

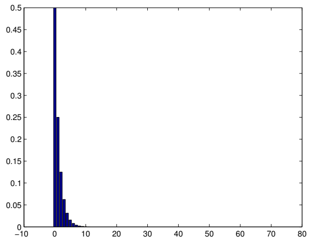
1.2.2 The Geometric Distribution

The Geometric random variable $Geo(p)$

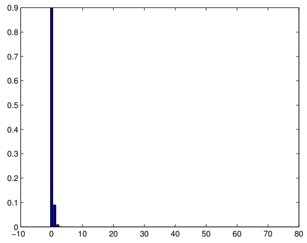
- Used to model the number of trials needed until a success occurs in independent Bernoulli experiments.
- Has p.m.f. defined by $P(X = i) = (1 - p)^{i-1}p, i = 1, 2, ...$
- Decays exponentially. Is the discrete version of the exponential random variable.
- Has mean $\frac{1}{p}$, variance $\frac{1-p}{p^2}$.



$Geom(.1)$



$Geom(.5)$



$Geom(.9)$

1.2.3 First Steps in Modeling

Let's look again at the example from Section 1.1. We determined there the *empirical* distribution of the number of hurricanes in a given year. Let's try to see if the data fit a discrete distribution we are now familiar with. As mentioned above, we'd certainly want the sample mean and the theoretical mean to match. We found the sample mean to be $\frac{602}{111}$. Let's use this as the parameter in the Poisson and geometric random variables to see how well the tables match:

For the Poisson r.v $X \sim Po(\frac{171}{64})$ and the geometric r.v $Y \sim Geo(\frac{64}{71})$, we get (all numbers rounded after 3 decimals):

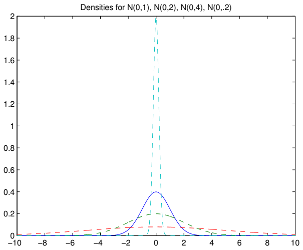
a	0	1	2	3	4	5	6	7	8
$P(X=a)$	0.069	0.185	0.247	0.220	0.147	0.078	0.035	0.013	0.004
$P(Y=a)$	0.374	0.234	0.147	0.092	0.057	0.036	0.022	0.014	0.009
%(games with a goals)	0.109	0.188	0.125	0.313	0.141	0.063	0.031	0.016	0.016

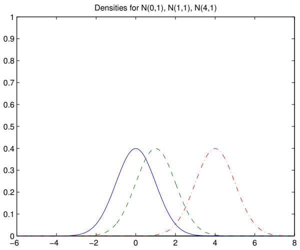
We see that the Poisson distribution is certainly better suited than the geometric to model these data.

1.2.4 The Normal Distribution

The Normal random variable $\mathcal{N}(\mu, \sigma^2), \mu, \sigma \in \mathbb{R}$

- Is used to model experiments which consist of sums of independent random experiments.
- Is therefore related to pretty much *every* random variable, through the Central Limit Theorem.
- Has density
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, x \in \mathbb{R}.$$
- Its mode occurs at $x = \mu$.
- The normal density is symmetric about the axis $x = \mu$.
- The inflection points of the normal density are at $x = \mu \pm \sigma$.
- Has mean μ, variance σ^2.





Definition 1.1. The *sample mean* of a family of random variables $\{X_1, \ldots, X_n\}$ is

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

The following is a very important property of the sample mean of i.i.d. normal random variables which we we will use repeatedly throughout the semester:

Proposition 1.1. Suppose $X_1, \ldots, X_n$ are i.i.d. $N(\mu, \sigma^2)$. Then

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Proof. Let $\Phi(a) = P(X_1 \leq a)$ be the c.d.f. of X_1. Then $\phi(a) := \frac{d}{da}\Phi(a)$ is the p.d.f. of X_1.

$$P(\frac{1}{n}X_1 \leq a) = P(X_1 \leq na) = \Phi(na).$$

So if f is the density of $\frac{1}{n}X_1$,

$$f(a) = \frac{d}{da}\Phi(na) = n\phi(na) = \frac{n}{\sqrt{2\pi}\sigma}e^{-\frac{(na-\mu)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi}\sigma/n}e^{-\frac{(a-\mu/n)^2}{2(\sigma/n)^2}},$$

which is the density of a $N(\mu/n, \sigma^2/n^2)$.

We know via moment generating functions that if $Y_i \sim N(\mu_i, \sigma_i^2)$ are independent, then $\sum_{i=1}^{n} Y_i \sim N(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2)$. So

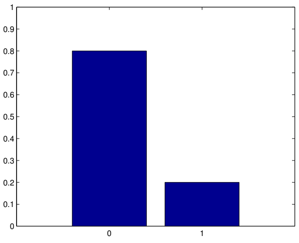
$$\frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu, \sigma^2/n).$$

□

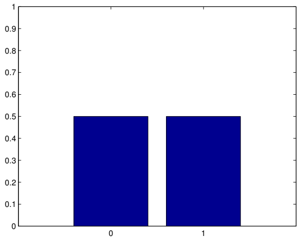
1.2.5 Other Distributions

The Bernoulli random variable $Be(p), 0 \leq p \leq 1$.

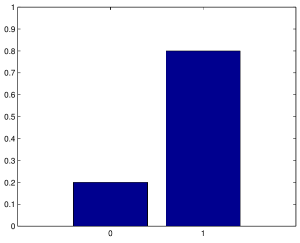
- Used to model an experiment which can only result in one of two outcomes (0 or 1).
- Has p.m.f. defined by $P(X = 1) = p, P(X = 0) = 1 - p$.
- Is connected to the binomial random variable as follows: If $X_i, i = 1, 2, ...$ are i.i.d. Bernoulli $Be(p)$, then $\sum_{i=1}^{n} X_i \sim Bin(n, p)$.
- Has mean p and variance $p(1 - p)$.



$Be(.2)$



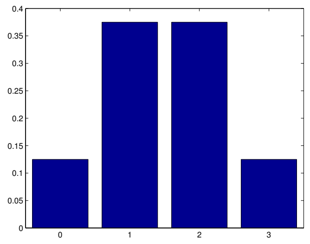
$Be(.5)$



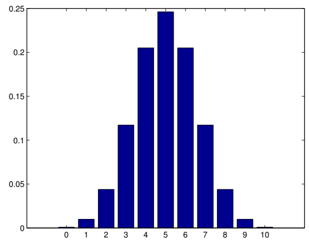
$Be(.8)$

The Binomial random variable $Bin(n,p), n \in \mathbb{N}, 0 \leq p \leq 1$

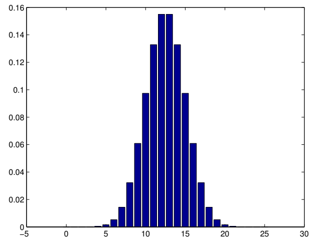
- Used to model the number of successes in n repeated independent identical Bernoulli experiments.
- Has p.m.f. defined by $P(X=i) = \binom{n}{i} p^i (1-p)^{n-i}, i = 0,..,n$.
- In particular, if $X \sim Bin(1,p)$, then $X \sim Be(p)$.
- Its mode occurs at the largest integer $k \leq (n+1)p$.
- Can be approximated by the Poisson random variable.
- Has mean np and variance $np(1-p)$ (which can be shown in one line using the fact that a binomial is a sum of independent Bernoullis).



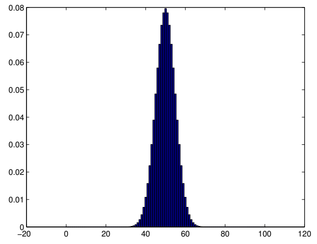
$Bin(3,.5)$



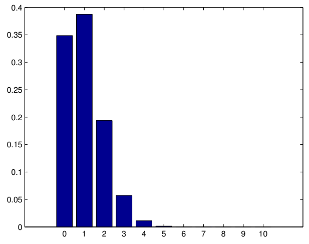
$Bin(10,.5)$



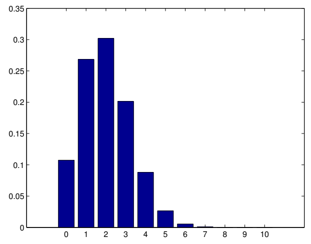
$Bin(25,.5)$



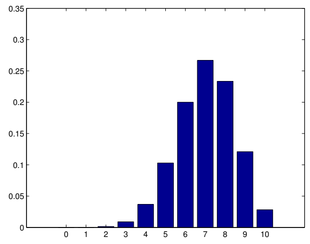
$Bin(100,.5)$



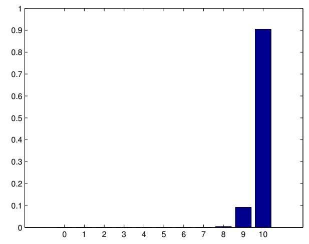
$Bin(10,.1)$



$Bin(10,.2)$



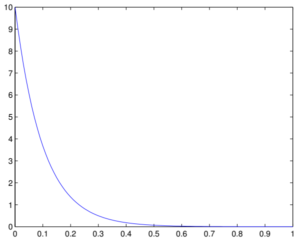
$Bin(10,.7)$



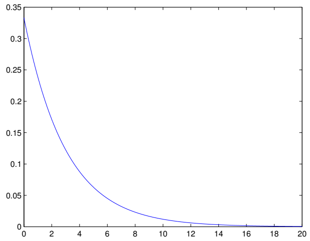
$Bin(10,.99)$

The Exponential random variable $Exp(\lambda), \lambda > 0$

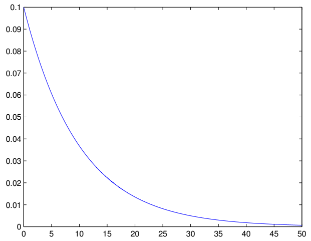
- Used to model the amount of time until an event occurs
- Has density
$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & otherwise \end{cases}$$
- Has the lack of memory property.
- Is related to the Gamma random variable as follows. $Exp(\lambda)$ has the same distribution as $\Gamma(1, \lambda)$. Moreover, if $X_i \sim Exp(\lambda), i = 1, 2, ...$ are i.i.d. , then $\sum_{i=1}^{n} X_i \sim \Gamma(n, \lambda)$.
- Has mean $\frac{1}{\lambda}$, variance $\frac{1}{\lambda^2}$.



$Exp(.1)$



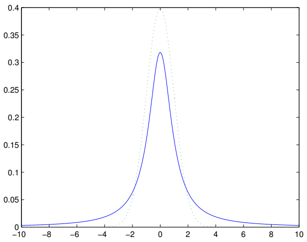
$Exp(3)$



$Exp(10)$

The Cauchy random variable

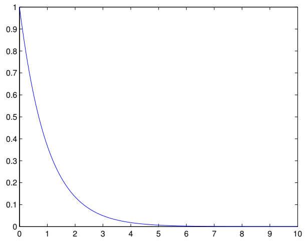
- Unlike most common random variables, the Cauchy has infinite variance.
- Has density
$$f(x) = \frac{1}{\pi(1 + x^2)}, \quad x \in \mathbb{R}$$
- Is symmetric about the y-axis.
- Is connected to the normal: If X and Y are independent standard normals, then $\frac{X}{Y}$ is Cauchy.
- If X is a Cauchy r.v., $\frac{1}{X}$ is Cauchy as well.
- Satisfies $E[|X|] = \infty, Var(X) = \infty$.



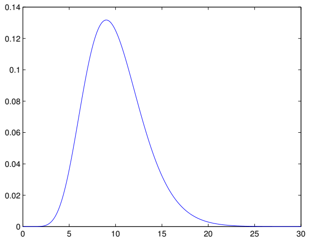
The Cauchy density drawn together with the standard normal. (Which is which?)

The Gamma random variable $\Gamma(\alpha, \lambda)$ $(\alpha > 0, \lambda > 0)$

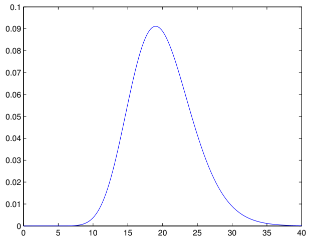
- When α is an integer, the Gamma r.v. is a sum of Exponential r.v.s, in which case it can be used to model the amount of time until α events occur.
- Has density
$$f(x) = \begin{cases} \frac{\lambda e^{-\lambda x}(\lambda x)^{\alpha-1}}{\Gamma(\alpha)}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$
- Has mean $\frac{\alpha}{\lambda}$, variance $\frac{\alpha}{\lambda^2}$.



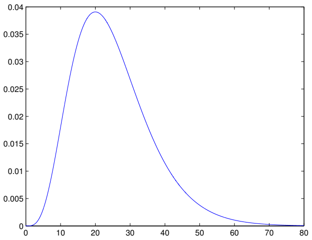
$\Gamma(1, 1)$ $(= Exp(1))$



$\Gamma(10, 1)$



$\Gamma(20, 1)$



$\Gamma(5, 5)$

Math 4501 (Spring 2018) January 31, 2018
Prof. Christian Beneš

Lecture #2: Order Statistics

2.1 Order Statistics

Definition 2.1. If $Y_1, \dots, Y_n$ are i.i.d. and all have the same distribution as Y, we define $Y_{(i)}$ to be the ith smallest value of $Y_1, \dots, Y_n$. We use the following, more explicit notation in two of the cases:

$$Y_{\min} := Y_{(1)}, \qquad Y_{\max} := Y_{(n)}.$$

$Y_{(i)}$ is called the *ith order statistic.*

Note 2.1. By definition,

$$Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n-1)} \leq Y_{(n)}.$$

It is important to note that the $Y_{(i)}$ are random variables too (since their values depend on the random values of $Y_1, \dots, Y_n$. We will eventually need to deal with the distributions of these random variables, so let's see how to obtain them.

We start with a simpler case, that of $Y_{\max} = \max\{\mathrm{Y}_1, \dots, \mathrm{Y}_\mathrm{n}\}$. You may remember that when looking for the p.d.f. of a random variable, it is often judicious to first look for the c.d.f. The present setting is no exception to that general rule.

$$\begin{aligned} F_{Y_{\max}}(y) &= P(Y_{\max} \leq y) = P(Y_1 \leq y, \dots, Y_n \leq y) \overset{indep.}{=} P(Y_1 \leq y) \cdots P(Y_n \leq y) \\ &\overset{i.d.}{=} (P(Y \leq y))^n = (F_Y(y))^n . \end{aligned}$$

Therefore,

$$f_{Y_{\max}}(y) = \frac{d}{dy} (F_Y(y))^n = n\, (F_Y(y))^{n-1} \frac{d}{dy} (F_Y(y)) = n\, (F_Y(y))^{n-1} f_Y(y).$$

Now let's turn to the general case. We will assume that Y is a continuous distribution, so that the probability that any of the Y_i are equal is 0. In order for for the jth order statistic to be equal to y, one of the r.v.'s $Y_1, \dots, Y_n$ has to be equal to y, $j-1$ of them have to be less than y, and $n-j$ of them have to be greater than y.

- The probability that in a given set of $j-1$ Y_i's, all Y_i's are less than y is $F_Y(y)^{j-1}$.
- The probability that in a given set of $n-j$ Y_i's, all Y_i's are greater than y is $(1 - F_Y(y))^{n-j}$.
- The probability that a given Y_i is in $[y, y+dy]$ is approximately $F_Y(y+dy) - F_Y(y)$.

The probability that all this happens simultaneously is (by independence) approximately

$$F_Y(y)^{j-1}(1-F_Y(y))^{n-j}(F_Y(y+dy)-F_Y(y)).$$

Now we also need to take into account the fact that this state of things (i.e., that $Y_{(j)}=y$) can be attained by a number of different configurations. The number of ways of splitting the n r.v.'s into three groups, one of $j-1$ elements, one of 1 element, and one of $n-j$ elements is

$$\binom{n}{j-1,1,n-j}.$$

Therefore,

$$P(Y_{(j)}\in[y,y+dy])\approx\binom{n}{j-1,1,n-j}F_Y(y)^{j-1}(1-F_Y(y))^{n-j}(F_Y(y+dy)-F_Y(y)),$$

so that

$$f_{Y_{(j)}}(y)\approx\frac{P(Y_{(j)}\in[y,y+dy])}{dy}\approx\binom{n}{j-1,1,n-j}F_Y(y)^{j-1}(1-F_Y(y))^{n-j}\frac{F_Y(y+dy)-F_Y(y)}{dy}.$$

When dy goes to 0, the " $\approx$ " above become " $=$ " and $\frac{F_Y(y+dy)-F_Y(y)}{dy}\to F'_Y(y)=f(y)$, so we get

$$f_{Y_{(j)}}(y)=\binom{n}{j-1,1,n-j}F_Y(y)^{j-1}(1-F_Y(y))^{n-j}f_Y(y).$$

Note, in particular that this implies that

$$f_{Y_{\max}}(y)=\binom{n}{n-1,1,0}F_Y(y)^{n-1}(1-F_Y(y))^0f_Y(y)=nF_Y(y)^{n-1}f_Y(y),$$

as we already showed, and that

$$f_{Y_{\min}}(y)=\binom{n}{0,1,n-1}F_Y(y)^0(1-F_Y(y))^{n-1}f_Y(y)=n(1-F_Y(y))^{n-1}f_Y(y).$$

Math 4501 (Spring 2018)
Prof. Christian Beneš

February 5, 2018

Lecture #3: Parameter Estimation

3.1 Maximum Likelihood Estimation

3.1.1 the one-parameter case

We start by answering two questions:

1. If $X \sim Po(\lambda)$ and λ is fixed, what is the value of k for which $P(X = k)$ is maximal?
2. If $X \sim Po(\lambda)$ and k is fixed, what is the value of λ for which $P(X = k)$ is maximal?

Although these questions look very similar, they are very different in nature. The first is a question a probabilist would ask, while the second is a statistician's question.

1. For $k \geq 1$,
$$\frac{P(X=k)}{P(X=k-1)} = \frac{\lambda}{k} \geq 1 \iff \lambda \geq k.$$
So if $\lambda \geq k, P(X = k) \geq P(X = k-1)$ and if $\lambda < k, P(X = k) < P(X = k-1)$.

   Therefore, the maximum will be reached at $k = [\lambda]$. Here, $[x]$ denotes the integer part of x.

2. Since λ is a continuous parameter, we can differentiate:
$$\frac{d}{d\lambda}\left(e^{-\lambda}\frac{\lambda^k}{k!}\right) = e^{-\lambda}\left(\frac{k\lambda^{k-1}}{k!} - \frac{\lambda^k}{k!}\right) = 0 \iff k - \lambda = 0 \iff k = \lambda.$$
Taking second derivatives, we see that $\lambda = k$ is a maximum.

A great place to experiment with this is:

http://www.causeweb.org/repository/statjava/PoiDensityApplet.html

Example 3.1. Suppose that a sample is drawn from a Poisson distribution and the outcome is 3. What is your best guess for λ? The computation above shows that the most likely value for λ is 3.

What if we draw again an independent sample from the same distribution and obtain the number 4? If we call X_1 and X_2 the two independent random variables of which we've observed the realization, we get

$$P(X_1 = 3, X_2 = 4) = e^{-\lambda}\frac{\lambda^3}{3!}e^{-\lambda}\frac{\lambda^4}{4!}.$$

Setting the derivative of the expression on the right to 0, we see that this is maximal when $\lambda = \frac{7}{2}$.

We can even take this one step further: Suppose $X_1, \ldots, X_n$ are i.i.d. samples from a Poisson random variable. What parameter λ makes the outcome

$$X_1 = x_1, \ldots, X_n = x_n$$

most likely?

$$L(\lambda) = P(X_1 = x_1, \ldots X_n = x_n) = e^{-\lambda}\frac{\lambda^{x_1}}{x_1!} \cdots e^{-\lambda}\frac{\lambda^{x_n}}{x_n!} = e^{-n\lambda}\frac{\lambda^{x_1+\ldots x_n}}{x_1! \ldots x_n!}.$$

This function is unfortunately not straightforward to maximize (in terms of λ), so we'll use a simple idea which will often tremendously simplify calculations for us: If f is maximal at x_0, $ln(f)$ is also maximal at x_0. In other words,

The functions $f(x)$ and $\ln(f(x))$ are maximal at the same point.

This is due to nothing else than the fact that the function $\ln(x)$ is monotonically increasing. More precisely, since $f(x) = e^{\ln(f(x))}$ and by the chain rule we have

$$\frac{d}{dx}(f(x)) = \frac{d}{dx}(e^{\ln(f(x))}) = e^{\ln(f(x))}\frac{d}{dx}(\ln(f(x))),$$

we see that, $\frac{d}{dx}(f(x)) = 0$ if and only if $\frac{d}{dx}(\ln(f(x))) = 0$ since $e^{\ln(f(x))} > 0$.

So instead of looking for the maximum of the function $L(\lambda)$ above, we can look for the maximum of $\ln(L(\lambda))$:

$$\ln(L(\lambda)) = \ln(e^{-n\lambda}\frac{\lambda^{x_1+\ldots x_n}}{x_1! \ldots x_n!}) = -n\lambda + (x_1 + \ldots x_n)\ln(\lambda) - \ln(x_1! \ldots x_n!).$$

So now, we can find the maximum of this function by setting its derivative to zero:

$$\frac{d}{d\lambda}\left(-n\lambda + (x_1 + \ldots x_n)\ln(\lambda) - \ln(x_1! \ldots x_n!)\right) = -n + \frac{x_1 + \ldots x_n}{\lambda} = 0 \iff \lambda = \frac{x_1 + \ldots x_n}{n}.$$

So the maximum of $L(\lambda)$ is reached at

$$\lambda = \frac{x_1 + \ldots x_n}{n} = \bar{x},$$

that is, if λ is the mean of the observed data.

The example above illustrates a more general method, that of maximum likelihood.

Definition 3.1. If $X_1 = x_1, \ldots, X_n = x_n$ is a random sample from the discrete p.m.f. $p_X(x;\theta)$ (respectively, continuous p.d.f. $f_X(x;\theta)$), the likelihood function $L(\theta)$ is the product of the p.m.f.'s (respectively, p.d.f.'s) evaluated at the k_i's, that is

$$L(\theta) = \prod_{i=1}^{n} p_X(x_i;\theta) \qquad \left(\text{respectively, } L(\theta) = \prod_{i=1}^{n} f_X(x_i;\theta)\right).$$

Definition 3.2. A function of a random sample whose aim is to approximate/estimate a parameter is called a *statistic* or an *estimator*. If θ is the parameter, the estimator is denoted by $\hat{\theta}$. The value found when evaluating the estimator is the *estimate*.

Note 3.1. The estimator is a random variable, the estimate a number.

Note 3.2. In the Poisson example above with $x_1 = 3, x_2 = 4$, the estimator was

$$\hat{\lambda} = \frac{X_1 + X_2}{2} = \bar{X},$$

while the estimate was

$$\lambda_e = \frac{x_1 + x_2}{2} = \frac{3+4}{2} = \frac{7}{2} = \bar{x}.$$

Definition 3.3. If $L(\theta)$ is the likelihood function corresponding to a random sample from a distribution and if θ_e is a value of the parameter such that $L(\theta_e) \geq L(\theta)$ for all θ, then θ_e is called the *maximum likelihood estimate* for θ. The (prior) function of the data yielding the maximum likelihood estimate is the *maximum likelihood estimator.*

Example 3.2. If

$$f(x) = \begin{cases} \frac{1}{\theta} & 0 < x < \theta \\ 0 & \text{otherwise} \end{cases},$$

find the maximum likelihood estimator $\hat{\theta}$ for θ.

Solution:

$$L(\theta) = \prod_{i=1}^{n} f_X(x_i; \theta),$$

where the x_i are the outcomes of the data. So

$$L(\theta) = \begin{cases} \frac{1}{\theta^n} & 0 < x_1, \ldots x_n < \theta \\ 0 & \text{otherwise} \end{cases}.$$

We usually like to find maxima by differentiating and setting the derivative equal to 0. The problem here is that there is no solution to $\frac{d}{d\theta} L(\theta) = 0$. We have to think about $L(\theta)$ a bit more carefully...

Since θ must be $\geq$ the largest value of the x_i, we have the constraint $\theta \geq \max_{1 \leq i \leq n} x_i$. On the other hand, since $L(\theta)$ is decreasing when it isn't 0, we want θ to be as small as possible. The smallest it can be all the while satisfying $\theta \geq \max_{1 \leq i \leq n} x_i$ is $\theta_e = \max_{1 \leq i \leq n} x_i = x_{\text{max}}$. In particular the maximum likelihood estimator is

$$\hat{\theta} = \max_{1 \leq i \leq n} X_i = X_{\text{max}}.$$

3.1.2 the multi-parameter case

When dealing with several parameters $\theta_1, \ldots, \theta_r$, we get a likelihood function $L(\theta_1, \ldots, \theta_r)$ which we want to maximize. As above, there are several ways of doing this, but if the function is differentiable, one way of maximizing it is by solving

$$\vec{\nabla} \ln L(\theta_1, \ldots, \theta_r) = \vec{0},$$

which is the same as solving

$$\frac{\partial}{\partial \theta_i} \ln L(\theta_1, \dots, \theta_r) = 0, \qquad 1 \le i \le r.$$

Example 3.3. This idea can be used to find the maximum likelihood estimator $(\hat{\mu}, \hat{\sigma^2})$ for a sample of i.i.d. normal random variables.

3.2 Method of Moments Estimation

The main idea is that if our model is good, the **theoretical moments** should be close to the **sample moments.** For this to make sense, we need to define sample moments. This is done in the natural and intuitive way by analogy with the definition of moments of a random variable.

Definition 3.4. If $x_1, \dots, x_n$ are random samples from the p.d.f $f_X(x; \theta_1, \dots, \theta_s)$, then

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

is called the jth *sample moment.*

Note 3.3. The first sample moment is just the sample mean. By the law of large numbers, the sample mean looks more and more like the true mean as n gets large. Similarly, the jth sample moment looks more and more like the true jth moment as n gets large.

Definition 3.5. The *method of moments estimates* $\theta_{1,e}, \dots, \theta_{s,e}$ for the parameters $\theta_1, \dots \theta_s$ in a distribution are the solutions of the set of equations

$$\frac{1}{n}\sum_{i=1}^{n} x_i = E[X]$$

$$\vdots$$

$$\frac{1}{n}\sum_{i=1}^{n} x_i^s = E[X^s],$$

where X is a random variable with the given distribution.

Note 3.4. In general, finding the method of moments estimate for a random variable depending on s variables amounts to solving s equations in s unknowns.

Example 3.4. Suppose $x_1 = 1.2, x_2 = 3, x_3 = 0.8$ are drawn from a gamma distribution with parameters r and λ. What is the method of moments estimate for (r, λ)?

Solution: First recall that if $X \sim \Gamma(r, \lambda)$, then $E[X] = r/\lambda$ and $Var(X) = r/\lambda^2$, so that $E[X^2] = Var(X) + E[X]^2 = \frac{r+r^2}{\lambda^2}$.

The first sample moment is

$$\frac{1}{3}\sum_{i=1}^{3} x_i = \frac{5}{3}$$

and the second sample moment is

$$\frac{1}{3}\sum_{i=1}^{3} x_i^2 = 3.69\bar{3}.$$

We therefore need to solve

$$\left\{ \begin{array}{ll} E[X] & = \frac{5}{3} \\ E[X^2] & = 3.69\bar{3} \end{array} \right. \Rightarrow \left\{ \begin{array}{ll} \frac{r}{\lambda} & = \frac{5}{3} \\ \frac{r+r^2}{\lambda^2} & = 3.69\bar{3} \end{array} \right. .$$

From the first equation, we get $r = (5/3)\lambda$, which, when plugged into the second equation gives

$$\frac{\frac{5}{3}\lambda + \left(\frac{5}{3}\lambda\right)^2}{\lambda^2} = 3.69\bar{3} \Rightarrow \left(\left(\frac{5}{3}\right)^2 - 3.69\bar{3}\right)\lambda = -\frac{5}{3} \iff \lambda = \frac{5/3}{0.91\bar{5}} \iff \lambda \approx 1.82.$$

Therefore, $r = \frac{5}{3}\lambda \approx 3.034$. All this gives

$$(r_e, \lambda_e) \approx (3.034, 1.82).$$

Math 4501 (Spring 2018) February 7, 2018
Prof. Christian Beneš

Lecture #4: Interval Estimation

Suppose we estimate a parameter θ of a distribution using some method (e.g. of moments or max. likelihood) and find a value θ_e. How comfortable should we feel that this is close to reality?

This of course depends on the distribution of $\hat{\theta}$.

Draw a few pictures with different variances.

In the second picture θ_e is more likely to be far away from the true value of θ.

Instead of focusing on finding a single value for θ, we will try to find intervals which are **very likely** to contain θ.

4.0.1 Confidence intervals for the mean of a normal distribution with known variance

Example 4.1. Suppose $X \sim N(\mu, 9)$ and if we draw four samples, we get $x_1 = 2.3, x_2 = 4.3, x_3 = 8.6, x_4 = 4.8$, how good an estimate is the MLE?

First, let's find the MLE: A few computational steps [to be filled in later] give $\hat{\mu} = \bar{X}$, so the maximum likelihood estimate is 5.

Now we know from Proposition 1.1 that $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$, which implies that in our case

$$\frac{\bar{X}-\mu}{3/\sqrt{4}} = \frac{2}{3}(\bar{X}-\mu) \sim N(0,1).$$

We will now set out to answer the following question: Can we find an interval which has a 95% chance of containing μ?

For $0 < \alpha \leq 1/2$, let's define z_α to be the number for which

$$P(Z \geq z_\alpha) = \alpha.$$

Using the normal table, we find that $z_{0.025} \approx 1.96$, which implies that

$$P(-1.96 \leq Z \leq 1.96) \approx 0.95.$$

Our computation above now implies

$$\begin{aligned} P(-1.96 \leq 2(\bar{X}-\mu)/3 \leq 1.96) \approx 0.95 \quad &\Rightarrow P(-2.94 \leq \bar{X}-\mu \leq 2.94) \approx 0.95 \\ &\Rightarrow P(\bar{X}-2.94 \leq \mu \leq \bar{X}+2.94) \approx 0.95. \end{aligned}$$

Therefore, the random interval $[\bar{X}-2.94, \bar{X}+2.94]$ will contain μ roughly 95% of the time (i.e., has probability 95% of containing the parameter μ). In our example, the intervals is approximately $[2.06, 7.94]$. It is called a *95% confidence interval* for μ.

In the example above, we can replace 0.05 by α and let the sample size be an arbitrary n, then based on n independent samples $X_i \sim N(\mu, \sigma^2)$, we can use the same procedure to find an interval which will have $100(1-\alpha)\%$ probability of containing μ.

$$P\left(-z_{\alpha/2} \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1-\alpha \Rightarrow P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha.$$

Therefore, an interval which has $100(1-\alpha)\%$ probability of containing μ is

$$\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right).$$

This interval is called a $100(1-\alpha)\%$ *confidence interval for* μ.

Note 4.1. The confidence interval we derived here is for the mean μ of a normal random variable with unknown mean μ and known mean σ^2 and is based on n independent samples from that random variable.

4.0.2 Confidence intervals for the binomial parameter p

When performing a poll where two options are given (Yes/No, Obama/McCain, etc.), one often assumes that voting intentions from one individual to the next. In this context, we are dealing with independent Bernoulli random variables, so that their sum is a binomial r.v.

Recall that if $X \sim Bin(n,p)$, we have $E[X] = np$ and $Var(X) = np(1-p)$, so that, by the central limit theorem (since a binomial is a sum of independent random variables, namely Bernoulli r.v.'s),

$$\frac{X-np}{\sqrt{np(1-p)}} \overset{\text{approx}}{\sim} N(0,1)$$

or, equivalently,

$$\frac{X/n-p}{\sqrt{p(1-p)/n}} \overset{\text{approx}}{\sim} N(0,1) \tag{1}$$

The goal is now to come up with a set of inequalities for p as we did above for μ. Based on (1), this would be difficult (if you're not convinced, try it), but we can approximate one level further and use the fact that (by the law of large numbers) X/n approximates p, so that we can rewrite (1) as

$$\frac{X/n-p}{\sqrt{\frac{X/n(1-X/n)}{n}}} \overset{\text{approx}}{\sim} N(0,1).$$

This implies that

$$P\left(-z_{\alpha/2} \leq \frac{X/n-p}{\sqrt{\frac{X/n(1-X/n)}{n}}} \leq z_{\alpha/2}\right) \approx 1-\alpha. \tag{2}$$

We can now use this to construct a confidence interval for p. We can do this exactly as in the previous subsection by putting p in the center of two inequalities, which will be based on (2). This gives the following *approximate* $100(1-\alpha)\%$ confidence interval for p:

$$\left(X/n - z_{\alpha/2}\sqrt{\frac{X/n(1-X/n)}{n}}, X/n + z_{\alpha/2}\sqrt{\frac{X/n(1-X/n)}{n}}\right).$$

Note 4.2. The approximate confidence interval we just derived relied heavily on the fact that a binomial random variable "looks enough like" a normal random variable. This is not going to be the case if n is small. We'll get back to what we mean by "n small" later.

Let's see how these ideas can be applied in the context of polls:

Example 4.2. In a survey conducted between January 12 and 14, 2007, 600 adults were asked the question: " Do you approve or disapprove the job Ted Kennedy is doing as a U.S. Senator?" The pollsters reported that 67% of the population approved Ted Kennedy's job, with a margin of error of 4%.

What does this mean? It turns out that we need a bit more information to understand the pollsters' statement precisely.

The actual data were ("data" is actually a plural, so since I'm a bit pedantic, I'll treat it as such even though most people don't) as follows: 402 of the polled people said they approve Kennedy's job, 198 said they don't. We can thus look for a confidence interval for the *true* value of the parameter p representing the probability that an individual from the entire population (not just the sampled people) approves Kennedy's job. Let's look (as is usually the case) for a 95% confidence interval. All that we have to do is to plug into the formula above 402 for X, 600 for n, and 1.96 for $z_{0.025}$. Then an approximate 95% confidence interval for p is

$$(0.67 - 0.038, 0.67 + 0.038) \approx (0.632, 0.711).$$

This means that with confidence of 95% (make sure you think about what this means), we can say that p is within 3.8% of 67%, So why did the pollsters say 4% when in fact they did a bit better than that? The reason is that they gave the *maximal margin of error*, a concept which we now discuss.

Definition 4.1. The *margin of error* is half of the width of the confidence interval (note that this depends on p). The *maximum margin of error* is half of the greatest possible width of the confidence interval (regardless of what p is).

We already know from our computations above that the margin of error is

$$z_{\alpha/2}\sqrt{\frac{X/n(1-X/n)}{n}}.$$

To find the maximum margin of error, we need to know how big $X/n(1-X/n)$ can be. Since $0 < X/n < 1$, we need to maximize the function $f(p) = p(1-p)$ for $0 \leq p \leq 1$. Since we're experts at calculus, this will be a piece of cake.

Lemma 4.1. If $0 \leq p \leq 1$,

$$p(1-p) \leq \frac{1}{4}.$$

Proof. Let $f(p) = p(1-p) = p - p^2$. Then $f'(p) = 1 - 2p = 0 \iff p = \frac{1}{2}$. We check the critical points and boundary points: $f(0) = f(1) = 0$; $f(1/2) = 1/4$. Therefore, f reaches its maximum of 1/4 at 1/2. □

A consequence of this lemma is that the maximum margin of error is

$$\frac{z_{\alpha/2}}{2\sqrt{n}}. \tag{3}$$

Note 4.3. Sometimes pollsters will report the maximum margin of error rather than the true (or *specific*) margin of error. This is what happened in the example above. Indeed, $\frac{1.96}{2\sqrt{600}} \approx 4\%$.

So why does one care about the maximum margin of error? Well, as you can see in the formula above, the max. margin of error depends on n only. This means that before conducting a poll, you can choose what you would like the margin of error to be at most and choose n accordingly. Here are some values for the maximum margin of error for different values of n and two values of α. They are obtained from (3) above.

sample size	at 95% confidence level (in %)	at 90% confidence level (in %)
60	12.7	10.6
100	9.8	8.2
300	5.7	4.7
500	4.4	3.7
800	3.5	2.9
1000	3.1	2.6

To know what n to choose given a desired maximum margin of error, we have the following easy result:

Theorem 4.1. If X/n is the estimator for p in a binomial distribution, in order for X/n to have at least a $100(1-\alpha)\%$ probability of being within distance d of p, the sample size should be at least

$$n = \frac{z_{\alpha/2}^2}{4d^2}.$$

Proof. We want to find the smallest n for which

$$1-\alpha = P(-d \leq X/n \leq d) \approx P\left(\frac{-d\sqrt{n}}{\sqrt{p(1-p)}} \leq Z \leq \frac{d\sqrt{n}}{\sqrt{p(1-p)}}\right).$$

But we know that (by definition)

$$1-\alpha = P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1-\alpha.$$

Therefore, by Lemma 4.1,

$$\frac{d\sqrt{n}}{\sqrt{p(1-p)}} = z_{\alpha/2} \Rightarrow n = \frac{z_{\alpha/2}^2 p(1-p)}{d^2} \leq \frac{z_{\alpha/2}^2}{4d^2}.$$

□

Math 4501 (Spring 2018) February 14, 2018
Prof. Christian Beneš

Lecture #5: Unbiasedness

5.1 Unbiased Estimators

One would hope that an estimator $\hat{\theta}$ for a parameter θ would, on average, yield the actual value of θ. This is addressed by the notion of unbiasedness.

Definition 5.1. If $Y_1, \ldots, Y_n$ is a random sample from some distribution $f(y;\theta)$, then $\hat{\theta}$ is *unbiased* if

$$E[\hat{\theta}] = \theta,$$

regardless of the actual value of θ. If $\hat{\theta}$ is not unbiased, it is *biased.*

Example 5.1. If $Y_1, \ldots, Y_n$ are drawn from a normal distribution with unknown mean μ, variance σ^2, then the ML and the method of moments estimators for μ and σ^2 are

$$\hat{\mu} = \bar{Y}, \quad \hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^n (Y_i - \bar{Y})^2.$$

Are any of these unbiased?

$$E[\hat{\mu}] = E[\bar{Y}] = E[\frac{1}{n}\sum_{i=1}^n Y_i] = \frac{1}{n}\sum_{i=1}^n E[Y_i] = \frac{1}{n}n\mu = \mu,$$

so $\hat{\mu}$ is unbiased.

Before computing $E[\hat{\sigma^2}]$, let's re-write $\hat{\sigma^2}$ in a more convenient way:

$$\begin{aligned}
\hat{\sigma^2} &= \frac{1}{n}\sum_{i=1}^n (Y_i - \bar{Y})^2 = \frac{1}{n}\sum_{i=1}^n (Y_i^2 - 2Y_i\bar{Y} + \bar{Y}^2) \\
&= \frac{1}{n}\left(\sum_{i=1}^n Y_i^2 - 2\sum_{i=1}^n Y_i\bar{Y} + \sum_{i=1}^n \bar{Y}^2\right) \\
&= \frac{1}{n}\left(\sum_{i=1}^n Y_i^2 - 2\bar{Y}\sum_{i=1}^n Y_i + n\bar{Y}^2\right) \\
&= \frac{1}{n}\left(\sum_{i=1}^n Y_i^2 - 2\bar{Y}n\bar{Y} + n\bar{Y}^2\right) \\
&= \frac{1}{n}\left(\sum_{i=1}^n Y_i^2 - n\bar{Y}^2\right)
\end{aligned}$$

Therefore, since $Y_i \sim N(\mu, \sigma^2)$, so that $E[Y_i^2] = \sigma^2 + \mu^2$ and $\bar{Y} \sim N(\mu, \sigma^2/n)$, so that $E[\bar{Y}^2] = \sigma^2/n + \mu^2$, we have

$$\begin{aligned} E[\hat{\sigma^2}] &= \frac{1}{n} E\left[\sum_{i=1}^n Y_i^2 - n\bar{Y}^2\right] = \frac{1}{n}\sum_{i=1}^n E\left[Y_i^2\right] - E[\bar{Y}^2] \\ &= \frac{1}{n} n(\sigma^2 + \mu^2) - (\sigma^2/n + \mu^2) = \sigma^2(1 - \frac{1}{n}) = \sigma^2 \frac{n-1}{n}. \end{aligned}$$

Therefore, $\hat{\sigma^2}$ is biased. Can we find an unbiased estimator for σ^2? Sure. Let $S^2 := \frac{n}{n-1}\hat{\sigma^2}$. Then

$$E[S^2] = \frac{n}{n-1} E[\hat{\sigma^2}] = \frac{n}{n-1}\sigma^2 \frac{n-1}{n} = \sigma^2,$$

so S^2 is unbiased.

Definition 5.2. The estimator $S^2 := \frac{1}{n-1}\sum_{i=1}^n (Y_i - \bar{Y})^2$ is the **sample variance**. $S = \sqrt{S^2}$ is the **sample standard deviation**.

Math 4501 (Spring 2018) February 20, 2018
Prof. Christian Beneš

Lecture #6: Efficiency; Minimum Variance Estimators; the Cramér-Rao Bound

6.1 Efficiency

Another desirable property of estimators is that they be likely to be close to the true value of the parameter, something that is not guaranteed entirely by the sole property of unbiasedness.

Definition 6.1. If $\hat{\theta}_1$ and $\hat{\theta}_2$ are two unbiased estimators for θ, $\hat{\theta}_1$ is **more efficient** than $\hat{\theta}_2$ if $\mathrm{Var}(\hat{\theta}_1) < \mathrm{Var}(\hat{\theta}_2)$. The **relative efficiency** of $\hat{\theta}_1$ with respect to $\hat{\theta}_2$ is $\frac{\mathrm{Var}(\hat{\theta}_2)}{\mathrm{Var}(\hat{\theta}_2)}$.

Example 6.1. Let $Y_1, \ldots, Y_n$ have p.d.f.

$$f_Y(y) = \begin{cases} 1/\theta, & 0 < y < \theta \\ 0, & \text{otherwise} \end{cases}.$$

Find an unbiased estimator for θ based on $Y_{\max}$. Is it more efficient than the method of moments estimator?

First, let's look for the method of moments estimator. $\bar{Y} = E[Y] \iff \bar{Y} = \theta/2 \iff \theta = 2\bar{Y}$ so the method of moments estimator is $\hat{\theta}_1 = 2\bar{Y}$.

Recall from a previous lecture that the maximum likelihood estimator for θ is $Y_{\max}$. Is it unbiased? To determine that we need its distribution.

Recall that

$$f_{Y_{\max}}(y) = n(F_Y(y))^{n-1} f_Y(y).$$

In our particular case,

$$F_Y(y) = \begin{cases} 0, & y < 0, \\ y/\theta, & 0 < y < \theta \\ 1, & y > \theta \end{cases},$$

so

$$f_{Y_{\max}}(y) = \begin{cases} 0, & y < 0, \\ n\left(\frac{y}{\theta}\right)^{n-1}\frac{1}{\theta}, & 0 < y < \theta \\ 0, & y > \theta \end{cases} = \begin{cases} 0, & y < 0, \\ n\frac{y^{n-1}}{\theta^n}, & 0 < y < \theta \\ 0, & y > \theta \end{cases}.$$

From this, we can compute

$$E[Y_{\max}] = \int_0^\theta n\frac{y^{n-1}}{\theta^n} y\,dy = \int_0^\theta n\frac{y^n}{\theta^n}\,dy = \frac{n}{n+1}\frac{y^{n+1}}{\theta^n}\Bigg|_0^\theta = \frac{n}{n+1}\theta.$$

Therefore, $\hat{\theta}_2 = \frac{n+1}{n} Y_{\max}$ is an unbiased estimator.

To determine which of $\hat{\theta}_1$ and $\hat{\theta}_2$ is more efficient, we need to compute their variances.

- Since $Y_1 \sim \mathcal{U}(0,\theta)$, we have $\mathrm{Var}(Y_1) = \theta^2/12$. Therefore,

$$\mathrm{Var}(\hat{\theta}_1) = \mathrm{Var}(2\bar{Y}) = 4\,\mathrm{Var}(\frac{1}{n}\sum_{i=1}^{n} Y_i) = \frac{4}{n^2}\sum_{i=1}^{n}\mathrm{Var}(Y_i) = \frac{4}{n^2}n\frac{\theta^2}{12} = \frac{\theta^2}{3n}.$$

- Note first that

$$E[Y_{\max}^2] = \int_0^\theta y^2\frac{n}{\theta}\left(\frac{y}{\theta}\right)^{n-1} dy = \frac{n}{\theta^n(n+2)}\theta^{n+2} = \frac{n}{n+2}\theta^2.$$

Therefore,

$$\mathrm{Var}(\hat{\theta}_2) = E[\hat{\theta}_2^2] - E[\hat{\theta}_2]^2 = \frac{(n+1)^2}{n(n+2)}\theta^2 - \theta^2 = \frac{n^2+2n+1-n^2-2n}{n(n+2)}\theta^2 = \frac{\theta^2}{n(n+2)}.$$

Which of $\hat{\theta}_1$ and $\hat{\theta}_2$ is more efficient depends on n:

$$n(n+2) > 3n \iff n^2 - n > 0 \iff n(n-1) > 0 \iff n < 0 \text{ or } n > 1.$$

Therefore, $\hat{\theta}_2$ is more efficient than $\hat{\theta}_1$ for all $n \geq 2$. Its relative efficiency with respect to $\hat{\theta}_1$ is $\frac{n(n+2)}{3n} = \frac{n+2}{3}$.

6.2 Minimum Variance Estimators

One might wonder if there is a limit to how efficient an unbiased estimator can be. The answer, perhaps not too surprisingly, is yes.

Note 6.1. The requirement of unbiasedness is of course important for this question to be interesting, as the estimator $\hat{\theta} = 5$ (for any parameter of any distribution) has zero variance, but is unbiased (and certainly a very bad estimator).

Theorem 6.1 (Cramér-Rao Bound)**.** Suppose $Y_1, \ldots, Y_n$ is a sample from the pdf $f_Y(y;\theta)$, where f_Y has continuous first and second order partial derivatives everywhere except perhaps at a finite number of points and the domain of f_Y doesn't depend on θ. Let $\hat{\theta}$ be an unbiased estimator for θ. Then

$$\mathrm{Var}(\hat{\theta}) \geq \left(nE\left[\left(\frac{\partial \ln f_Y(Y;\theta)}{\partial\theta}\right)^2\right]\right)^{-1} = \left(-nE\left[\frac{\partial^2 \ln f_Y(Y;\theta)}{\partial\theta^2}\right]\right)^{-1}.$$

The same inequality holds for discrete distributions such that the p.m.f. doesn't depend on θ.

Definition 6.2. Let Θ be the set of all unbiased estimators $\hat{\theta}$ for θ. We say that $\theta^* \in \Theta$ is a **best** or **minimum variance unbiased estimator (MVUE)** if $\mathrm{Var}(\theta^*) \leq \mathrm{Var}(\hat{\theta})$ for all $\hat{\theta} \in \Theta$. An unbiased estimator is **efficient** if its variance equals the Cramér-Rao lower bound.

Note 6.2. If an unbiased estimator is efficient, then it is a MVUE. The other implication does not hold.

Example 6.2 (Cautionary tale)**.** If $Y_1, \dots, Y_n \sim \mathcal{U}(0, \theta)$, then we know from earlier that $\hat{\theta} = \frac{n+1}{n} Y_{\max}$ is an unbiased estimator and $\mathrm{Var}(\hat{\theta}) = \frac{\theta^2}{n(n+2)}$. Also,

$$\begin{aligned}\left(nE\left[\left(\frac{\partial \ln f_Y(y;\theta)}{\partial \theta}\right)^2\right]\right)^{-1} &= \left(nE\left[\left(\frac{\partial}{\partial \theta}\ln(1/\theta)\right)^2\right]\right)^{-1} \\ &= \left(nE\left[-\frac{1}{\theta^2}\right]\right)^{-1} = \frac{\theta^2}{n}.\end{aligned}$$

But this implies that for all $n, \mathrm{Var}(\hat{\theta})$ is smaller than the Cramér-Rao lower bound! The problem here is that we were not allowed to use the theorem, since the domain of f_Y depends on the parameter θ.

Example 6.3. Suppose $X_1, \dots, X_n \sim Po(\lambda)$, that is, $P_X(k;\lambda) = \frac{e^{-\lambda}\lambda^k}{k!}$. Then $\hat{\lambda} = \bar{X}$ is an unbiased estimator for λ (check this!). Is $\hat{\lambda}$ an MVUE?

Note first that

$$\mathrm{Var}(\hat{\lambda}) = \mathrm{Var}(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{1}{n^2}n\lambda = \frac{\lambda}{n}.$$

Let's compute the Cramér-Rao bound:

$$\frac{\partial}{\partial \lambda}\ln p_X(X;\lambda) = \frac{\partial}{\partial \lambda}(-\lambda + X\ln(\lambda) - \ln(k!)) = -1 + \frac{X}{\lambda}.$$

Therefore,

$$E\left[\left(\frac{\partial}{\partial \lambda}\ln p_X(X;\lambda)\right)^2\right] = E\left[1 - \frac{2X}{\lambda} + \frac{X^2}{\lambda^2}\right] = 1 - \frac{2\lambda}{\lambda} + \frac{\lambda + \lambda^2}{\lambda^2} = \frac{1}{\lambda},$$

and so

$$\left(nE\left[\left(\frac{\partial}{\partial \lambda}\ln p_X(X;\lambda)\right)^2\right]\right)^{-1} = \frac{\lambda}{n},$$

which implies that $\hat{\lambda}$ is efficient and therefore an MVUE.

Math 4501 (Spring 2018) February 21, 2018
Prof. Christian Beneš

Lecture #7: Consistency; sufficiency

7.1 Consistency

Definition 7.1. An estimator $\hat{\theta}_n = f(Y_1, \ldots, Y_n)$ is **consistent** if for every $\epsilon > 0$,

$$\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| > \epsilon) = 0.$$

Draw a picture.

Example 7.1. Suppose $X \sim N(\mu, \sigma^2)$ and $X_1, \ldots, X_n$ are independent samples with the same distribution as X. Let $\hat{\mu}_n^1 = X_1$ and $\hat{\mu}_n^1 = \bar{X}_n$. Which of $\hat{\mu}_n^1$ and $\hat{\mu}_n^2$ are consistent?

First note that $E[\hat{\mu}_n^1] = \mu$, so $\hat{\mu}_n^1$ is unbiased. However,

$$P(|\hat{\mu}_n^1 - \mu| > \epsilon) = P(|X_1 - \mu| > \epsilon) = 2\int_{-\infty}^{\mu-\epsilon} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx,$$

which is a constant and therefore can't go to 0.

As we already know, $\hat{\mu}_n^2$ is unbiased. Moreover, since $\bar{X}_n \sim N(\mu, \sigma^2/n)$, we have, by Chebyshev's inequality,

$$P(|\hat{\mu}_n^2 - \mu| > \epsilon) = P(|\bar{X}_n - \mu| > \epsilon) \le \frac{\text{Var}(\bar{X}_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \to 0, \text{ as } n \to \infty.$$

Note 7.1. We can see from the argument of the last example that in order to show consistency of an unbiased estimator $\hat{\theta}_n$, it suffices to show that $\text{Var}(\hat{\theta}_n) \to 0$ as $n \to \infty$.

7.2 Sufficiency

The concept of a *sufficient statistic* is a little bit more difficult to grasp well than the other properties of statistics we've discussed so far. In a few words, a statistic is sufficient for a given parameter if knowing its value is just as good as knowing the whole sample for the purpose of making the estimate. For example, if one tries to estimate θ based on a sample $X_1, \ldots, X_n \sim \mathcal{U}(0, \theta)$, we've seen before that the two estimators for θ, $\hat{\theta}_1 = 2\bar{X}$ and $\hat{\theta}_2 = X_{\max}$ are, respectively, the method of moments and maximum likelihood estimators for θ. Intuitively, are they sufficient to estimate the parameter θ or would knowing the whole sample $X_1, \ldots, X_n$ have any additional influence on the choice of θ? For example

- if I see the entire data set and tell you that $x_{\max} = 5$, might I come to a better conclusion than you (i.e., that $\theta = 5$) by seeing the entire data set?
- if I see the entire data set and tell you that $2\bar{x} = 4$, might I come to a better conclusion than you (i.e., that $\theta = 4$) by seeing the entire data set?

The answer in the first case is no, while in the second it is yes. Indeed, suppose that the data set was $1, 2, 1, 5, 1$. Then $\bar{x} = 2$, so that the method of moments estimate is 4. However, having seen the entire data set, I can tell you that this estimate is not one that I would make. Indeed, having seen the value 5, I know that there is no way that θ could be 4. In this case, just knowing the value of $\bar{x}$ was not *sufficient* to estimate θ well.

The formal definition of sufficiency is as follows:

Definition 7.2. Let $Y_1, \ldots, Y_n$ be a random sample from a probability distribution with unknown parameter θ. The statistic $U = g(Y_1, \ldots, Y_n)$ is said to be *sufficient* for θ if the conditional distribution of $Y_1, \ldots, Y_n$ given U doesn't depend on θ.

Since conditional distributions are not always easy to compute, the following *factorization theorem* will come in very handy:

Theorem 7.1. (Factorization Theorem) Let U be a statistic based on the random sample $Y_1, \ldots, Y_n$. Then U is a sufficient statistic for θ if and only if the likelihood $L(\theta)$ can be factored into two nonnegative functions

$$L(\theta) = g(u, \theta) \cdot h(y_1, \ldots, y_n),$$

where $g(u, \theta)$ is a function only of u and θ and $h(y_1, \ldots, y_n)$ is not a function of θ.

Example 7.2. Let's revisit the motivating example from earlier. Suppose $Y_1, \ldots, Y_n \sim \mathcal{U}(0, \theta)$. Then as we have seen already,

$$L(\theta) = \frac{1}{\theta^n} \mathbb{1}\{0 \leq \theta \leq y_{\max}\} \cdot 1,$$

so using the factorization theorem with $u = y_{\max}$, $g(u, \theta) = \frac{1}{\theta^n} \mathbb{1}\{0 \leq \theta \leq Y_{\max}\}$, $h(y_1, \ldots, y_n) = 1$, we see that $U = Y_{\max}$ is a sufficient statistic.

Example 7.3. Find a sufficient estimator for p when $X_1, \ldots, X_n \sim \mathit{Geo}(p)$.

$$L(\theta) = \prod_{i=1}^n P(X_i = x_i) = \prod_{i=1}^n (1-p)^{x_i} p = (1-p)^{\sum_{i=1}^n x_i} \left(\frac{p}{1-p}\right)^n = g(h(x_1, \ldots, x_n), p) \cdot 1,$$

where $b(x_1, \ldots, x_n) = 1, h(x_1, \ldots, x_n) = \sum_{i=1}^n x_i$ and $g(a, p) = (1-p)^a \left(\frac{p}{1-p}\right)^n$, so by the factorization theorem, $U = \sum_{i=1}^n X_i$ is a sufficient statistic.

Math 4501 (Spring 2018) February 26, 2018
Prof. Christian Beneš

Lecture #8: Hypothesis Testing

8.1 Introduction

Example 8.1. A colony of laboratory mice consists of several thousand mice. The average weight of all the mice is 32g. with standard deviation 4g. A lab assistant was asked by a scientist to select 25 mice for an experiment. However, before performing the experiment, the scientist decided to weigh the mice as an indicator of whether the assistant's selection constituted a random sample or whether it was made with some unconscious bias. If the sample mean of the mice was 30.4, would this be significant evidence (at the 5% level of significance) against the hypothesis that the selection constituted a random sample? Assume the weights of the mice in the colony are normally distributed.

We'll return to the example shortly, but first a few generalities about hypothesis testing:

Hypothesis testing works like the judicial system: Innocent unless proven guilty.

Step 1: Set up a null hypothesis (H_0) and an alternative hypothesis (H_1). Usually H_0 represents status quo, that is, the situation where nothing unusual is happening.

In our case,

$$\begin{aligned} H_0 &: \text{The sample is random (i.e., the assistant did do his job well)} \\ H_1 &: \text{There is a bias in the sample (i.e., the assistant didn't do his job well)} \end{aligned}$$

We want to check if there is a strong case against H_0. We assume that 32 is the true mean weight of the mice. In mathematical terms, if μ is the mean weight of the mice selected by the assistant, we want to test $H_0 : \mu = 32$ against $H_1 : \mu \neq 32$.

What is the probability that something as extreme as or more than what we observed would happen under the assumption H_0? Here $X_1, \dots, X_{25}$ are the random weights of the 25 selected mice. Then under H_0, $\bar{X} \sim N(32, \frac{16}{25})$.

$$\begin{aligned} &P(\text{something as extreme as or more than what we observed would happen under the assumption } H_0) \\ &\quad = P(\bar{X} \leq 30.4 \text{ or } \bar{X} \geq 33.6 | H_0 \text{ true}) \\ &\quad = P(|\bar{X} - 32| \geq 1.6 | H_0 \text{ true}) = 2P(\bar{X} \leq 30.4 | H_0 \text{ true}) \\ &\quad = 2P\left(\frac{\bar{X} - 32}{4/5} \leq \frac{30.4 - 32}{4/5} | H_0 \text{ true}\right) = 2P(Z \leq -2) \approx 0.0455 \end{aligned}$$

The probability we just computed is the *p-value.* It measures the probability that something as extreme as or more than what we observed would happen under the assumption H_0? Usually, we will reject H_0 only if the p-value is small. In general, we reject H_0 at the α significance level if the p-value is less than or equal to α. In our particular example, we'd reject H_0 at the 5% level of significance.

Another equivalent way of testing such a hypothesis is by setting up a *critical region*, the region in which the statistic (in this case $\bar{X}$) must fall in order for H_0 to be rejected.

In that case, one determines beforehand that

$$P(\text{we reject } H_0 | H_0 \text{ is true}) = \alpha,$$

In our example, this would lead to the equation

$$\begin{aligned} P(|\bar{X} - 32| \geq a | H_0 \text{ true}) = 5\% &\iff 2P(\bar{X} \geq a + 32 | H_0 \text{ true}) = 5\% \\ &\iff 2P(\frac{\bar{X} - 32}{4/5} \geq \frac{a}{4/5} | H_0 \text{ true}) = 5\% \\ &\iff P(Z \geq 5a/4) = 0.025. \end{aligned}$$

Therefore, using values from the normal table we find

$$\frac{5a}{4} \approx 1.96 \iff a \approx 1.56.$$

So we should reject H_0 if and only if

$$|\bar{X} - 32| \geq 1.56 \iff \bar{X} \geq 33.56 \text{ or } \bar{X} \leq 30.44.$$

Here are the general definitions related to hypothesis testing:

Definition 8.1. A function of the observed data whose numerical value dictates whether H_0 is rejected or not is called a **test statistic**. The values of the test statistic for which H_0 is rejected are the **critical region**. The number(s) on the boundary between the critical region is (are) called the **critical value(s)**. The probability α that the test statistic lies in the critical region if H_0 is true is the **level of significance** of the test. The $p-$**value** associated with a test statistic is the probability is the probability that we get a value as extreme as or more extreme than what was observed, relatively to H_1, given that H_0 is true.

Note 8.1. α is a number you choose *a priori* (before conducting the experiment). p is a number you obtain *a posteriori* (after the experiment has been conducted).

If the p-value is less than α, you will reject H_0 at the α level of significance.

The critical region is determined both by α and by the alternative hypothesis H_1. We will see how this works in the particular case of testing for the mean of a normal random variable with known variance:

Testing $H_0 : \mu = \mu_0$ for $Y_1, \ldots Y_n \sim N(\mu, \sigma^2)$ with σ known

We let $z = \frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}}$

1. (one-sided test) To test $H_0 : \mu = \mu_0$ against $H_1 : \mu > \mu_0$ (respectively, $H_1 : \mu < \mu_0$) at the α level of significance, we reject H_0 if $z \geq z_\alpha$ (respectively, $z \leq -z_\alpha$).
2. (two-sided test) To test $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$, we reject H_0 if $z \geq z_{\alpha/2}$ or $z \leq -z_{\alpha/2}$.

Note 8.2. One should use 2-sided test if there is no *a priori* reason for suspecting that $\mu > \mu_0$ rather than $\mu < \mu_0$. For example, suppose we are testing whether steroids increase strength. If μ_0 is the mean weight people can lift, then for a sample of people on steroids, one would test $H_0 : \mu = \mu_0$ against $H_1 : \mu > \mu_0$.

Math 4501 (Spring 2018) March 6, 2017
Prof. Christian Beneš

Lecture #9: Type I and Type II Errors; Binomial Data

9.1 Type I and Type II Errors

When setting up hypothesis tests, we usually choose a level of significance α which yields a critical region for the statistic/estimator we are measuring. If it falls within the critical region, we reject H_0. This is our decision rule. Of course, our decision can be wrong. One of two bad things can happen:

- H_0 is actually true and we reject it. This is a Type I error.
- H_0 is false and we fail to reject it. This is a Type II error.

We can put the 4 possible configurations in a table:

	H_0 true	H_0 false
Fail to reject H_0	✓	Type II Error
Reject H_0	Type I Error	✓

By definition of the level of significance,

$$P(\text{Type I error}) = P(\text{Reject } H_0 | H_0 \text{ is true}) = \alpha.$$

Moreover,

$$P(\text{Type II error}) = P(\text{Fail to reject } H_0 | H_0 \text{ is false}).$$

H_0 being false typically encompasses a large number of scenarios which need to be considered separately.

Example 9.1. Suppose that $X_1, \dots, X_{120} \sim \mathcal{U}(0, \theta)$ and we wish to test $H_0 : \theta = 4$ against $H_1 : \theta \neq 4$ using the maximum likelihood estimator $\hat{\theta} = X_{\max}$. Suppose we have chosen the decision rule to be that we fail to reject H_0 if $3.9 \leq X_{\max} \leq 4$ and reject otherwise. Then

$$\begin{aligned}
\alpha &= P(\text{type I error}) = P(\text{reject } H_0 | H_0 \text{ true}) \\
&= P(X_{\max} < 3.9 | \theta = 4) + P(X_{\max} > 4 | \theta = 4) \\
&= P(X_{\max} < 3.9 | \theta = 4) = \int_0^{3.9} \frac{120}{4} \left(\frac{y}{4}\right)^{n-1} dy = \left(\frac{3.9}{4}\right)^{120} \approx 0.048.
\end{aligned}$$

Also,

$$\beta = P(\text{type II error}) = P(\text{fail to reject } H_0 | H_0 \text{ false}).$$

There are uncountably many ways in which H_0 can be false, so we'll compute

$$P(\text{fail to reject } H_0 | \theta) = P(3.9 \leq X_{\max} \leq 4 | \theta)$$

for all θ (in particular the cases of interest are $\theta \neq 4$). We subdivide this calculation into three cases:

- If $\theta \leq 3.9$,
$$P(3.9 \leq X_{\max} \leq 4|\theta) = 0.$$

- If $3.9 \leq \theta \leq 4$,
$$P(3.9 \leq X_{\max} \leq 4|\theta) = \int_{3.9}^{\theta} \frac{120}{\theta} \left(\frac{y}{\theta}\right)^{n-1} dy = \left(\frac{y}{\theta}\right)^{120} \Big|_{3.9}^{\theta} = \frac{\theta^{120} - 3.9^{120}}{\theta^{120}}.$$

- If $\theta \geq 4$,
$$P(3.9 \leq X_{\max} \leq 4|\theta) = \int_{3.9}^{4} \frac{120}{\theta} \left(\frac{y}{\theta}\right)^{n-1} dy = \left(\frac{y}{\theta}\right)^{120} \Big|_{3.9}^{4} = \frac{4^{120} - 3.9^{120}}{\theta^{120}}.$$

Graphs of θ vs β and θ vs $1-\beta$.

Definition 9.1. $1-\beta$ is the **power** of the test. The graph of θ versus $1-\beta$ is the **power curve** of the test.

Note 9.1. Increasing α decreases β and vice-versa. Increasing n helps both α and β (but costs more). Typically we care more about α than β.

Example 9.2. Suppose $H_0 : \mu = 10$ is tested against $H_1 : \mu < 10$ at the $\alpha = 0.01$ level of significance, based on $X_1, \ldots, X_n \sim N(\mu, 9)$. What is the smallest value of n for which the power of the test will be at least 0.75 if $\mu = 9$? Perform a test using the estimator $\bar{X}$ for μ.

We first need to determine the decision rule:

$$\begin{aligned} P(\text{reject} H_0 | H_0 \text{true}) = 0.01 \quad &\iff \quad P(\bar{X} \leq a | \mu = 10) = 0.01 \\ &\iff \quad P\left(\frac{\bar{X} - 10}{\sqrt{9/n}} \leq \frac{a - 10}{\sqrt{9/n}} \middle| \mu = 10\right) = 0.01 \\ &\iff \quad P(Z \leq \frac{a-10}{\sqrt{9/n}}) = 0.01. \end{aligned}$$

But we know that $P(Z \leq -2.33) = 0.01$, so we need to solve

$$-2.33 = \frac{a-10}{3}\sqrt{n} \iff a = -\frac{6.99}{\sqrt{n}} + 10.$$

So the rule requires that we reject H_0 if $\bar{X} \leq 10 - \frac{6.99}{\sqrt{n}}$.

Now we wish to find n such that

$$P(\text{fail to reject} H_0 | \mu = 9) = \beta \leq 0.25.$$

$$\begin{aligned} 0.25 \quad &\geq \quad P(\bar{X} \geq 10 - \frac{6.99}{\sqrt{n}} | \mu = 9) \overset{\bar{X} \sim N(9, 9/n)}{=} P\left(\frac{\bar{X} - 9}{3}\sqrt{n} \geq \frac{10 - \frac{6.99}{\sqrt{n}} - 9}{3}\sqrt{n}\right) \\ &= \quad P(Z \geq \frac{\sqrt{n}}{3} - 2.33) \end{aligned}$$

But we know that $P(Z \geq 0.67) \approx 0.25$, so we solve

$$\frac{\sqrt{n}}{3} - 2.33 \geq 0.67 \iff \sqrt{n} \geq 9 \iff n \geq 81.$$

Note that this corresponds to $a = 9.22$.

We can draw a picture with two normals with mean 9 and 10 for the case $n = 81$ where we see the probabilities α and β determined by areas under the curves to either side of 9.22.

Math 4501 (Spring 2018) March 5, 2018
Prof. Christian Beneš

Lecture #10: Testing Binomial Data - $H_0 : p = p_0$; Generalized Likelihood Ratio Test

10.1 Large-sample case

We will use the normal approximation to the binomial whenever we can. However, this is not always a reasonable thing to do.

Suppose $X \sim Bin(100, \frac{1}{100})$, so that $P(X = k) = \binom{100}{k} \left(\frac{1}{100}\right)^k \left(\frac{99}{100}\right)^{100-k}$. Then the graph of the p.m.f. of X doesn't look normal at all.

The problem is that the mean of X is too close to one of the extreme values of the distribution ("close" here is in the sense of number of standard deviations). If we were to normalise this binomial, we'd define

$$Z = \frac{X - np}{\sqrt{np(1-p)}} = \frac{X - 1}{\sqrt{99/100}},$$

a random variable that cannot take any values below $-\sqrt{\frac{99}{100}} \approx -1$.

Similarly, if $Y \sim Bin(100, \frac{99}{100})$, the random variable

$$Z = \frac{Y - np}{\sqrt{np(1-p)}} = \frac{Y - 99}{\sqrt{99/100}}$$

is a random variable that cannot take any values above $\sqrt{\frac{100}{99}} \approx 1$.

Some useful applets to play with this:

`http://digitalfirst.bfwpub.com/stats_applet/stats_applet_2_cltbinom.html`

`https://www.geogebra.org/m/CmHJuJxs`

If we require, somewhat arbitrarily, that the mean of the binomial be at least three standard deviations away from its extreme values (i.e., 0 and n), we don't have this problem anymore. This requirement is equivalent to

$$3\sqrt{np(1-p)} \leq np \leq n - 3\sqrt{np(1-p)}.$$

When this requirement is satisfied, we can perform a so-called **large-sample** hypothesis test.

Example 10.1. You are gambling with a friend. You win \$1 if the flip of a coin gives heads and lose \$1 if it gives tails. You suspect your friend's coin is not fair and that heads is more likely to come up. After 100 flips, you owe your friend \$20. Does a 5% significance hypothesis test confirm your doubts?

We let $p = P(\text{Tails comes up})$ and test

$H_0 : p = 1/2$,

$H_1 : p > 1/2$.

Let's check if the large-sample requirement is satisfied:

$$3\sqrt{100\frac{1}{2}\frac{1}{2}} = 15 \leq 50 \leq 100 - 15 = n - 3\sqrt{npq},$$

so we can perform a large-sample test.

The idea is that since the normal is a good approximation for the binomial. Let X be the number of tails. Then the outcome of the experiment was $X = 60$ (resulting in you winning \$40 and your friend \$60 for a net loss of \$20). Let's compute the p-value:

$$p = P(X \geq 60|H_0) = P\left(\frac{X-50}{5} \geq \frac{X-50}{5}\bigg|H_0\right) \approx P(Z \geq 2) \approx 0.023 < \alpha.$$

Therefore, we reject H_0. The difference between X and $100p$ is statistically significant.

Note 10.1. If we had been testing $H_0 : p = 1/2$ against $H_1 : p \neq 1/2$, we would have

$$P(X \geq 60|H_0) = 0.023 < 0.025 = \alpha/2.$$

This is closer, though we would still have rejected H_0. Having the option of a one-sided alternative, generally gives you a better chance to reject H_0.

See Theorem 6.3.1 for the formal procedure of the binomial large-sample test.

Example 10.2. (continued) What is the probability of a Type II error if $H_0 : p = 1/2, H_1 : p = 11/20$?

$$P(\text{fail to reject } H_0|H_0 \text{ false}) = P(\text{fail to reject } H_0|p = 11/20).$$

The critical region in the one-sided case is determined by the value K for which $P(X \geq K|p = 1/2) = 0.05$.

$$P(X \geq K|p = 1/2) = 0.05 \iff P\left(\frac{X-50}{5} \geq \frac{K-50}{5}\right) = 0.05 \iff \frac{K-50}{5} \approx 1.64 \iff K \approx 58.2,$$

so we reject H_0 if $X \geq 58.2$. Therefore,

$$P(\text{Type II Error}) = P(X \leq 58.2|p = 11/20) = P\left(\frac{X-55}{10\sqrt{\frac{11}{20}\frac{9}{20}}} \leq \frac{58.2-55}{10\sqrt{\frac{11}{20}\frac{9}{20}}}\right) \approx P(Z \leq 0.643) \approx 0.74.$$

Picture showing these probabilities.

10.2 Small-sample case

If the large-sample inequalities don't hold, we can usually deal with the binomial distribution directly.

Example 10.3. Can people tell the difference between Miller and Löwenbräu?

18 people were asked to taste 3 beers (2 Miller and one Löwenbräu) and determine which one was different. 8 people guessed correctly.

Let p be the probability that an individual guessed correctly. Then we are testing

$H_0 : p = 1/3$

$H_1 : p \neq 1/3$

Let X be the number of correct guesses and $\alpha = 0.05$.

Then we have

$$p - \text{value} = P(X \geq 8|p = \frac{1}{3}) \approx 0.108 \not< \alpha,$$

so we fail to reject H_0. We don't have enough evidence to say that people can tell the difference between the beers.

Alternatively (and more complicatedly), let's first find the critical region:

$$p(X \geq K|H_0) = \sum_{i=k}^{18} \binom{18}{i} (1/3)^i (2/3)^{18-i} \overset{!}{\leq} 0.05.$$

- $P(X = 18) = (1/3)^{18}$
- $P(X = 17) = 18(1/3)^{17}(2/3)$
- $P(X = 16) = \frac{18 \cdot 17}{2}(1/3)^{16}(2/3)^2 \approx 1.58 \cdot 10^{-6}$
- ...
- $P(X \geq 9) \approx 0.043$
- $P(X \geq 8) \approx 0.108$

These calculations show we should reject H_0 if $X \geq 9$.

The p-value given by the experiment is 0.108. Since this is greater than α we fail to reject (as already implied by the previous line).

10.3 The GLR

All the tests we have developed are based on distributions involving the parameter(s) for which we would like to make some inference. It may require some cleverness to come up with the right distributions. There are also cases in which we have more than one reasonable estimator for a given parameter and therefore more than one decision rule (e.g. for the uniform parameter $\theta, X_{\max}$ and $2\bar{X}$ are distinct estimators for θ with distinct distributions).

We now examine a systematic way of developing a test for a given parameter.

When testing hypotheses, we usually assume under H_0 that the parameter θ comes from some set ω. We also denote by Ω the set of all possible values of θ. The **generalized likelihood ratio** is

$$\Lambda = \frac{\max_\omega L(\theta; X_1, \dots, X_n)}{\max_\Omega L(\theta; X_1, \dots, X_n)}.$$

Note that θ can be a vector.

If the data are about as likely under H_0 as they would be under any assumption on the parameter, then H_0 explains the data well. It should be rejected if the numerator of Λ does a much worse job at explaining L than the denominator does. We will reject H_0 at the significance level α if λ, the realization of Λ is below some threshold λ^* defined by

$$P(0 < \Lambda \le \lambda^* | H_0 \text{ true}) = \alpha.$$

This test is the *Generalized Likelihood Ratio Test (GLRT).*

Example 10.4. Suppose $X_1, \dots, X_n \sim \mathcal{U}(0, \theta)$ and we wish to test $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$.

Since ω consists of only one value of θ, i.e., θ_0, we see that

$$\max_\omega L(\theta; X_1, \dots, X_n) = L(\theta_0; X_1, \dots, X_n) = \frac{1}{\theta_0^n}$$

since

$$L(\theta_0) = \begin{cases} (1/\theta_0)^n, & X_{\max} \le \theta_0 \\ 0 & \text{otherwise} \end{cases}.$$

Similarly, as we derived early this semester,

$$\max_\Omega L(\theta; X_1, \dots, X_n) = \frac{1}{X_{\max}^n}$$

since

$$L(\theta) = \begin{cases} (1/\theta)^n, & X_{\max} \le \theta \\ 0 & \text{otherwise} \end{cases}.$$

Therefore,

$$\Lambda = \frac{X_{\max}^n}{\theta_0^n}.$$

Now

$$\begin{aligned} \alpha &= P(0 < \Lambda \le \lambda^* | H_0 \text{ true}) = P(0 < \frac{X_{\max}^n}{\theta_0^n} \le \lambda^* | H_0 \text{ true}) \\ &= P(0 < X_{\max}^n \le \theta_0^n \lambda^* | H_0 \text{ true}) = P(0 < X_{\max} \le \theta_0 (\lambda^*)^{1/n} | H_0 \text{ true}). \end{aligned}$$

Now we know from earlier in the semester that the pdf of $X_{\max}$ in this setting is

$$f_{X_{\max}}(x) = \begin{cases} \dfrac{n}{\theta^n} x^{n-1}, & 0 \le x \le \theta \\ 0 & \text{otherwise} \end{cases}.$$

If H_0 is true, then

$$f_{X_{\max}}(x) = \begin{cases} \dfrac{n}{\theta_0^n} x^{n-1}, & 0 \leq x \leq \theta_0 \\ 0 & \text{otherwise} \end{cases},$$

so that

$$\alpha = P(0 < X_{\max} \leq \theta_0 (\lambda^*)^{1/n} | H_0 \text{ true}) = \int_0^{\theta_0(\lambda^*)^{1/n}} \frac{n}{\theta_0^n} x^{n-1}\, dx = \frac{(\theta_0(\lambda^*)^{1/n})^n}{\theta_0^n} = \lambda^*.$$

Therefore, the decision rule is to reject H_0 if $\lambda \leq \alpha$, i.e., when $x_{\max}^n \leq \alpha\theta_0^n$, i.e., when

$$x_{\max} \leq \theta_0 \alpha^{1/n}.$$

Note that this test was already used implicitly in an example Lecture 9 in the particular case where $n = 120$ and $\theta_0 = 4$. You were told that the decision rule is to reject if $X_{max} \leq 3.9$ and found $\alpha = 0.048$. And indeed, with these values

$$\theta_0 \alpha^{1/n} = 4 \cdot 0.048 \approx 3.9 = \lambda^*.$$

We will find the GLRT for the parameter μ in the case $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ with both parameters unknown, when testing $H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$.

Example 10.5. We have

$$\omega = \{(\mu, \sigma^2) : \mu = \mu_0, \sigma^2 \in \mathbb{R}_+\}$$

and

$$\Omega = \{(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_+\}$$

Since

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

which is maximized on ω_0 (as can be verified by setting the partial derivatives equal to 0) at the point $(\mu_0, \frac{1}{n}\sum(x_i - \mu_0)^2)$ and on Ω_0 at the point $(\bar{x}, \frac{1}{n}\sum(x_i - \bar{x})^2)$.

We then can use direct substitution to compute λ.

Math 4501 (Spring 2018) March 7, 2018
Prof. Christian Beneš

Lecture #11: Neyman-Pearson Lemma; $\frac{\bar{Y}-\mu}{S/\sqrt{n}}$

11.1 The Neyman-Pearson Lemma

Suppose we are testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_a$. We have

$$\text{power}(\theta_a) = P(\text{reject} H_0 | \theta = \theta_a).$$

What does the ideal power curve look like? DRAW IT (0 at θ_0 and 1 everywhere else). Every different test yields a different power curve, so we should choose the test for which the power curve looks as close to this ideal curve as possible. Fortunately, there is a result that tells us what that test is!

Definition 11.1. A hypothesis is *simple* if it uniquely specifies the distribution of the population from which the sample is taken. A hypothesis that is not simple is *composite*.

Example 11.1. 1. If $Y_1, \ldots, Y_n \sim N(\mu, 9)$, then $H_0 : \mu = 3$ is a simple hypothesis (since under it we know exactly how Y_i is distributed).

2. If $Y_1, \ldots, Y_n \sim N(\mu, \sigma^2)$ with σ^2 unknown, then $H_0 : \mu = 3$ is a composite hypothesis (since under it we don't know the exact distribution of Y_i).

Lemma 11.1. (Neyman-Pearson Lemma) Consider all α level hypothesis tests with simple null and alternative hypotheses. Then the likelihood ratio test with $H_0 : \theta = \theta_0, H_1 : \theta = \theta_1$ which rejects H_0 if

$$\Lambda(X_1, \ldots, X_n) = \frac{L(\theta_0; X_1, \ldots, X_n)}{L(\theta_1; X_1, \ldots, X_n)} \leq \lambda^*,$$

where λ^* is such that $P(\Lambda(X_1, \ldots, X_n) \leq \lambda^* | H_0) = \alpha$, is the *most powerful* test, i.e., the test with the highest power,

Example 11.2. Suppose $X_1, \ldots, X_n \sim N(\mu, 1)$ is a random sample of size $n = 100$. Find the most powerful test with $\alpha = 0.05$ for $H_0 : \mu = 0$ against $H_1 : \mu = 1$.

Since both hypotheses are simple, we can use the Neyman-Pearson Lemma. Note that

$$\begin{aligned}
\frac{L(0; X_1, \ldots, X_n)}{L(1; X_1, \ldots, X_n)} \leq \lambda^* \quad &\Longleftrightarrow \quad \frac{\frac{e^{-\sum_{i=1}^{100} X_i^2/2}}{\sqrt{2\pi}^n}}{\frac{e^{-\sum_{i=1}^{100}(X_i-1)^2/2}}{\sqrt{2\pi}^n}} \leq \lambda^* \\
&\Longleftrightarrow \quad e^{-\sum_{i=1}^{100}(X_i^2-(X_i-1)^2)/2} \leq \lambda^* \\
&\Longleftrightarrow \quad \sum_{i=1}^{100}(X_i - 1/2) \geq -\ln(\lambda^*) \Longleftrightarrow \bar{X} \geq \frac{1}{2} - 100\ln(\lambda^*)
\end{aligned}$$

So we reject H_0 if $\bar{X} \geq k$ for some constant k. Since

$$\begin{aligned}
0.05 &= P(\bar{X} \geq k | H_0) = P(\bar{X}\sqrt{100} \geq \sqrt{100}k) | H_0) \\
&= P(Z \geq 10k)
\end{aligned}$$

and $0.05 = P(Z \geq 1.645)$, we have $k = 0.1645$, so the power of the test when $\mu = 1$ is

$$P(\bar{X} \geq 0.1645|\mu = 1) = P(10(\bar{X} - 1) \geq -8.355|\mu = 1) = P(Z \geq -8.355) \approx 1.$$

This is the greatest power any such test can achieve.

11.2 Determining the Distribution of $\frac{\bar{Y}-\mu}{S/\sqrt{n}}$

We know that if $Y_1, \dots, Y_n \sim N(\mu, \sigma^2)$, then $\frac{\bar{Y}-\mu}{\sigma/\sqrt{n}} \sim N(0, 1)$. We can use this to make inferences about μ if σ is known. But in general, we don't know σ, although we have an estimator for it:

$$\hat{\sigma} = S_n = \frac{1}{\sqrt{n-1}} \sqrt{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}.$$

It is natural to wonder if replacing σ by S changes the distribution. The answer, provided by William Gossett, is that it does, very much so if n is small.

We will now go through a sequence of steps in order to determine the distribution of $\frac{\bar{Y}-\mu}{S/\sqrt{n}}$.

Definition 11.2. If $Z_1, \dots, Z_n$ are independent standard normal random variables,

$$U = \sum_{i=1}^{n} Z_i^2$$

has a *chi-squared distribution* with n degrees of freedom. We write $U \sim \chi_n^2$.

Recall that $X \sim \Gamma(r, \lambda)$ has the gamma distribution with parameters r and λ if

$$f_X(x) = \begin{cases} \frac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}.$$

We have $E[X] = r/\lambda$ and $\mathrm{Var}(X) = r/\lambda^2$.

Theorem 11.1. If $U \sim \chi_n^2$, then $U \sim \Gamma\left(\frac{n}{2}, \frac{1}{2}\right)$. That is, the chi-square distribution with n degrees of freedom is the same as the gamma distribution with parameters $n/2$ and $1/2$.

Proof. We'll first find the distribution of Z^2 when $Z \sim N(0, 1)$: For $a \geq 0$,

$$F_{Z^2}(a) = P(Z^2 \leq a) = P(-\sqrt{a} \leq Z \leq \sqrt{a}) = \Phi(\sqrt{a}) - \Phi(-\sqrt{a}),$$

where Φ is the standard normal c.d.f. Therefore, for $a > 0$,

$$f_{Z^2}(a) = \frac{d}{da} F_{Z^2}(a) = \frac{d}{da}\left(\Phi(\sqrt{a}) - \Phi(-\sqrt{a})\right) = \frac{1}{2\sqrt{a}}\phi(\sqrt{a}) + \frac{1}{2\sqrt{a}}\phi(-\sqrt{a}) = \frac{1}{\sqrt{2\pi}} a^{-1/2} e^{-a/2}.$$

Using the fact that $\Gamma(1/2) = \sqrt{\pi}$, we recognize this as the pdf of the $\Gamma(1/2, 1/2)$ distribution, so $Z^2 \sim \Gamma(1/2, 1/2)$.

We know that if for $i = 1, \ldots, n, X_i \sim \gamma(r_i, \lambda)$, then $\sum_{i=1}^n X_i \sim \Gamma(\sum_{i=1}^n r_i, \lambda)$. Therefore, since if $Z_i \sim N(0,1), Z_i^2 \sim \Gamma(1/2, 1/2)$, so that if $U \sim \chi_n^2$,

$$U = \sum_{i=1}^n Z_i^2 \sim \Gamma(n/2, 1/2).$$

□

Definition 11.3. If $Z \sim N(0,1)$ and $U \sim \chi_n^2$ are independent, then

$$T_n = \frac{Z}{\sqrt{U/n}}$$

is a t random variable with n degrees of freedom. We write $T_n \sim t_n$.

Lemma 11.2. If $T_n \sim t_n, T_n$ is symmetric.

Proof.

$$-T_n = \frac{-Z}{\sqrt{U}/n},$$

so since $-Z \sim N(0,1)$, $-T_n \sim t_n$. □

Theorem 11.2.

$$f_{T_n}(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)\left(1 + \frac{t^2}{n}\right)^{(n+1)/2}}, t \in \mathbb{R}.$$

Note 11.1. f_{T_n} is even, which confirms that T_n is symmetric.

Theorem 11.3. If $Y_1, \ldots, Y_n \sim N(\mu, \sigma^2)$ are independent, then

1. S^2 and $\bar{Y}$ are independent.
2. $\frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^n (Y_i - \bar{Y})^2 \sim \chi_{n-1}^2$.

Theorem 11.4. If $Y_1, \ldots, Y_n \sim N(\mu, \sigma^2)$ are independent, then

$$\frac{\bar{Y} - \mu}{S/\sqrt{n}} \sim t_{n-1}.$$

Proof.

$$\frac{\bar{Y} - \mu}{S/\sqrt{n}} = \frac{\frac{\bar{Y}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{(n-1)\sigma^2}}}.$$

By part 2. of the last theorem, $\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$ and is independent of $\frac{\bar{Y}-\mu}{\sigma/\sqrt{n}}$ and $\frac{\bar{Y}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$, so the theorem follows from the definition of the t_{n-1} distribution. □

Example 11.3. (Use of the chi-square distribution) A small variation of lactic acid concentration corresponds to a small variation in the taste of cheese (which is desirable). Experiments show that for n chunks of cheese in which the concentration of lactic acid is modeled by $X_i \sim N(\mu, 0.09)$,

$$Y = \frac{1}{n}\sum_{i=1}^{n} |X_i - \mu|^2 \geq 0.09,$$

a good measure of how concentrations differ from μ, leads to decreased sales, so producers will want to avoid this. For 10 chunks of cheese in which the concentration of lactic acid is modeled by $X_i \sim N(\mu, 0.09)$, what is $P(Y \geq 0.09)$?

$$P(Y \geq 0.09) = P\left(\frac{0.09}{10}\sum_{i=1}^{10} Z_i^2 \geq 0.09\right) = P(\sum_{i=1}^{10} Z_i^2 \geq 10) = P(X \geq 10) \approx 0.45,$$

where $X \sim \chi^2_n$. To compute the probability in question, one option is to use the following applet, also useful for the t and normal distributions: `https://surfstat.anu.edu.au/surfstat-home/tables/chi.php`

Math 4501 (Spring 2018) March 12, 2017
Prof. Christian Beneš

Lecture #12: Normal Data - Inference for μ and σ^2

We are assuming throughout this lecture that $Y_1, \ldots, Y_n \sim N(\mu, \sigma^2)$ are independent.

12.1 Inference for μ if σ^2 is unknown

The key idea here will be that $\frac{\bar{Y}-\mu}{S/\sqrt{n}} \sim t_{n-1}$.
To construct confidence intervals, the procedure will be the same as before except we'll be working with $\frac{\bar{Y}-\mu}{S/\sqrt{n}} \sim t_{n-1}$ rather than $\frac{\bar{Y}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$.

12.1.1 Confidence Intervals

We define $t_{\alpha,n}$ as follows:

Definition 12.1. Suppose that $T \sim t_n$. Then

$$P(T \geq t_{\alpha,n}) = \alpha.$$

Therefore, using the fact that $\frac{\bar{Y}-\mu}{S/\sqrt{n}} \sim t_{n-1}$, we have

$$P\left(-t_{\alpha/2,n-1} \leq \frac{\bar{Y}-\mu}{S/\sqrt{n}} \leq t_{\alpha/2,n-1}\right) = 1-\alpha.$$

So

$$P\left(-t_{\alpha/2,n-1}\frac{S}{\sqrt{n}} \leq \bar{Y}-\mu \leq t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\right) = 1-\alpha.$$

Therefore,

$$P\left(-t_{\alpha/2,n-1}\frac{S}{\sqrt{n}} - \bar{Y} \leq -\mu \leq t_{\alpha/2,n-1}\frac{S}{\sqrt{n}} - \bar{Y}\right) = 1-\alpha,$$

so

$$P\left(\bar{Y} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}} \leq \mu \leq \bar{Y} + t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\right) = 1-\alpha.$$

Therefore, the $(1-\alpha)100\%$ confidence interval for μ is

$$\left(\bar{Y} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}, \bar{Y} + t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\right).$$

Note 12.1. For comparison,

- if σ is assumed to be known, a $100(1-\alpha)\%$ confidence interval for μ is

$$\left(\bar{Y} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{Y} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right).$$

- If $Y_1, \ldots, Y_n \sim Be(p)$ are independent, an approximate $100(1-\alpha)\%$ confidence interval for p is (with $Y = \sum_{i=1}^n Y_i$)

$$\left(\bar{Y} - z_{\alpha/2} \frac{\sqrt{\frac{Y}{n}\frac{Y-k}{n}}}{\sqrt{n}}, \bar{Y} + z_{\alpha/2} \frac{\sqrt{\frac{Y}{n}\frac{Y-k}{n}}}{\sqrt{n}} \right).$$

12.1.2 Testing $H_0 : \mu = \mu_0$

Example 12.1. Automobile producers need to destroy some of their cars to test their solidity. They have built a specific car with the goal in mind that a certain type of frontal car crash would yield a damage whose cost is normally distributed with mean \$15,000. They crashed 15 cars and found, for the cost, a sample mean of \$15,789 and a standard deviation of \$4782.50. With μ the mean cost of the damage in dollars, test at the 5% significance level $H_0 : \mu = 15,000$ against $H_1 : \mu \neq 15,000$.

We compute the p-value: For $T_{14} \sim t_{14}$,

$$\begin{aligned}
P(\bar{X} \geq 15,789 \text{ or } \bar{X} \leq 14,221 | H_0 \text{ true }) &= 2P(\bar{X} \geq 15,789 | H_0 \text{ true }) \\
&= 2P(\frac{\bar{X} - 15,000}{4782.5/\sqrt{15}} \geq \frac{15,789 - 15,000}{4782.5/\sqrt{15}} | H_0 \text{ true }) \\
&= 2P(T_{14} \geq \frac{789}{4782.5/\sqrt{15}}) \\
&= 2P(T_{14} \geq 0.639) \approx 0.5331.
\end{aligned}$$

Therefore, we fail to reject H_0. There isn't strong evidence that the mean cost of the given frontal crash is not \$15,000.

12.2 Inference for σ^2

The key idea is to use S^2 as an estimator for σ^2. We will need the following definition:

Definition 12.2. Suppose that $X \sim \chi^2_n$. Then

$$P(X \leq \chi^2_{\alpha,n}) = \alpha.$$

The definition, together with the fact that $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$ implies that

$$P\left(\chi^2_{\alpha/2,n-1} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{1-\alpha/2,n} \right) = 1 - \alpha.$$

Therefore,

$$P\left(\frac{1}{\chi^2_{\alpha/2,n-1}} \geq \frac{\sigma^2}{(n-1)S^2} \geq \frac{1}{\chi^2_{1-\alpha/2,n}} \right) = 1 - \alpha,$$

so

$$P\left(\frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}\right) = 1-\alpha,$$

which implies that a $(1-\alpha)100\%$ confidence interval for σ^2 is

$$\left(\frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n}}, \frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}\right),$$

and a $(1-\alpha)100\%$ confidence interval for σ is

$$\left(\sqrt{\frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n}}}, \sqrt{\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}}\right).$$

Example 12.2. The National Center for Educational Statistics surveyed college graduates about the time needed to complete their Bachelor"s degree. Polling 101 people gave a sample mean of 5.15 years and a sample standard deviation of 1.68 years. Find a 99% confidence interval for the standard deviation of the time needed by college students to graduate.

Using a chi square applet or table gives $\chi^2_{199/200,100} \approx 140.17$ and $\chi^2_{1/200,100} \approx 67.33$. This gives the following 99% confidence interval:

$$I \approx \left(\frac{10 \cdot 1.68}{\sqrt{140}}, \frac{10 \cdot 1.68}{\sqrt{67.33}}\right) \approx (1.419, 2.047).$$

Math 4501 (Spring 2018) March 14, 2018
Prof. Christian Beneš

Lecture #13: Exam 1

Math 4501 (Spring 2018) March 19, 2018
Prof. Christian Beneš

Lecture #14: Equality of Means - Motivation

Example 14.1. (Random Dot Stereograms) If you know what is hidden behind a random dot stereogram, do you see it faster?

In an experiment, subjects were divided into two groups, one of which (NV) was given no information about the image hidden in a random dot stereogram, while the other (VV) was given verbal and visual information. The times (in seconds) needed for subjects to recognize the image are represented by $X_1, \ldots, X_{43}$ for the NV group and $Y_1, \ldots, Y_{35}$ for the VV group. The data are available at

`http://userhome.brooklyn.cuny.edu/cbenes/RDS.txt`

To load that data set into R, we use

```
> www=“http://userhome.brooklyn.cuny.edu/cbenes/RDS.txt”
> TIMES=read.table(www,fill=TRUE)
```

The following commands allow us to separate the data into the two groups :

```
> T1=TIMES[1:43,1]
> T2=TIMES[44:78,1]
```

Note that “[1:43,1]” means that we are considering rows 1 to 43 and the first column.

Typing “T1” or “T2” now allows us to see the data sets separately. We can also obtain the mean and standard deviation by using the commands “mean” and “sd”.

```
> mean(T1)
[1] 8.560465
> mean(T2)
[1] 5.551429
> sd(T1)
[1] 8.085412
> sd(T2)
[1] 4.801739
```

Math 4501 (Spring 2018) March 21, 2018
Prof. Christian Beneš

Lecture #15: Equality of Means - the Two-Sample t Test

We are assuming throughout this lecture that $X_1, \dots, X_n \sim N(\mu_X, \sigma^2), Y_1, \dots, Y_m \sim N(\mu_Y, \sigma^2)$ are independent. Note that we are assuming a same variance for data from both samples, but not a same mean.

15.1 Inference for $\mu_X - \mu_Y$

Definition 15.1. For random variables $X_1, \dots, X_n, Y_1, \dots, Y_m$, the *pooled variance* is

$$S_p^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2} = \frac{1}{n+m-2}\left(\sum_{i=1}^n (X_i - \bar{X})^2 + \sum_{i=1}^m (Y_i - \bar{Y})^2\right).$$

Theorem 15.1. If $X_1, \dots, X_n \sim N(\mu_X, \sigma^2), Y_1, \dots, Y_m \sim N(\mu_Y, \sigma^2)$ are independent, then

$$\frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{n+m-2}.$$

Proof. We have

$$T = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}} = \frac{\frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}}}{\sqrt{\frac{1}{n+m-2}\left(\frac{(n-1)S_X^2}{\sigma^2} + \frac{(m-1)S_Y^2}{\sigma^2}\right)}}. \tag{4}$$

Since $\bar{X} \sim N(\mu_X, \sigma^2/n)$ and $\bar{Y} \sim N(\mu_Y, \sigma^2/m)$ are independent, we know that $\bar{X} - \bar{Y}$ is normal. Note that

- $E[\bar{X} - \bar{Y}] = \mu_X - \mu_Y$
- $\mathrm{Var}(\bar{X} - \bar{Y}) \overset{\text{indep.}}{=} \mathrm{Var}(\bar{X}) + \mathrm{Var}(-\bar{Y}) = \mathrm{Var}(\bar{X}) + \mathrm{Var}(\bar{Y}) = \frac{1}{n} + \frac{1}{m}$.

This implies that for the numerator in the last expression of (4), we have

$$\frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim N(0, 1).$$

To determine the distribution of the denominator, we note that

$$\frac{(n-1)S_X^2}{\sigma^2} \sim \chi^2_{n-1} \text{ and } \frac{(m-1)S_Y^2}{\sigma^2} \sim \chi^2_{m-1} \text{ are independent,}$$

so that, since a sum of gammas is gamma and the chi square r.v. is a particular case of the gamma,

$$U = \frac{(n-1)S_X^2}{\sigma^2} + \frac{(m-1)S_Y^2}{\sigma^2} \sim \chi^2_{n+m-2}.$$

Moreover, for the same reason as in the one-sample case, the numerator and the denominator are independent, so we have

$$T = \frac{Z}{\sqrt{\frac{U}{n+m-2}}},$$

with $Z \sim N(0,1)$ and χ^2_{n+m-2} independent, so $T \sim t_{n+m-2}$. □

The theorem we just proved allows us to test equality of means $H_0 : \mu_X = \mu_Y$ since under the assumption H_0,

$$\frac{\bar{X} - \bar{Y}}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{n+m-2}.$$

Example 15.1. (Random Dot Stereograms) If you know what is hidden behind a random dot stereogram, do you see it faster?

In an experiment, subjects were divided into two groups, one of which (NV) was given no information about the image hidden in a random dot stereogram, while the other (VV) was given verbal and visual information. The times (in seconds) needed for subjects to recognize the image are represented by $X_1, \ldots, X_{43}$ for the NV group and $Y_1, \ldots, Y_{35}$ for the VV group. The data are available at

`http://userhome.brooklyn.cuny.edu/cbenes/RDS.txt`

(and originally at `http://lib.stat.cmu.edu/DASL/Datafiles/FusionTime.html`)

To load that data set into R, we use

```
> www=”http://userhome.brooklyn.cuny.edu/cbenes/RDS.txt”
> TIMES=read.table(www,fill=TRUE)
```

We separate the data into the two groups:

```
> T1=TIMES[1:43,1]
> T2=TIMES[44:78,1]
> hist(T1)
> hist(T2)
> mean(T1)
[1] 8.560465
> mean(T2)
[1] 5.551429
> sd(T1)
[1] 8.085412
```

```
> sd(T2)
[1] 4.801739
```

This is translated into the following statistics:

$$\bar{X} \approx 8.56, \quad S_X \approx 8.09, \quad S_X^2 \approx 65.45,$$

$$\bar{Y} \approx 5.55, \quad S_Y \approx 4.80, \quad S_Y^2 \approx 23.04.$$

We will test $H_0 : \mu_X = \mu_Y$ against $H_1 : \mu_X > \mu_Y$ at the 5% significance level. In order to compute the p-value, it will be useful to have

$$S_p^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2} \approx \frac{42 \cdot 65.45 + 34 \cdot 23.04}{76} \approx 46.48,$$

so that $S_p \approx 6.82$. Moreover, $\sqrt{\frac{1}{n} + \frac{1}{m}} = \sqrt{\frac{1}{43} + \frac{1}{35}} \approx 0.23$. Letting $T_{76} \sim t_{76}$, all this gives the following $p - value$:

$$P(\bar{X} - \bar{Y} > 8.56 - 5.55 | H_0) = P\left(\frac{\bar{X} - \bar{Y}}{1.55} > \frac{8.56 - 5.55}{1.55} | H_0\right) \approx P(T_{76} > 1.94) \approx 0.028,$$

so we reject H_0, which means that there is strong evidence suggesting that having verbal and visual information about the image in a random dot stereogram, increases the speed at which one sees it.

Note 15.1.

1. Was it OK to assume that the data are normal? Not quite. For one, the data are discrete, but more importantly, we'll see later that a test for normality suggests that we can't make that assumption.

   However, this is not a big problem, as t-tests are "robust" with respect to departure from normality: If Y_i are not normal, the distribution of $\frac{\bar{Y}-\mu}{\sigma/\sqrt{n}}$ is relatively unaffected by the actual distribution of the Y_i, as long as that distribution is not too skewed and n is not too small.

2. Was it OK to assume $\sigma_X^2 = \sigma_Y^2$? We'll see in our next class how to answer this question. You'll also see on the homework in Problem 9.2.15 how to address the case where $\sigma_X^2 \neq \sigma_Y^2$ are both unknown.

3. How do we test $H_0 : \mu_X = \mu_Y$ in the (unrealistic) case we know both σ_X^2 and σ_Y^2? Then

$$\frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \sim N(0,1).$$

Math 4501 (Spring 2018) March 26, 2018
Prof. Christian Beneš

Lecture #16: Equality of Variances - the F test

16.1 The F distribution

In order to test equality of variances, we need to introduce a new distribution.

Definition 16.1. Suppose $U \sim \chi^2_n$ and $V \sim \chi^2_m$ are independent. Then we define the F distribution with m and n degrees of freedom to be the distribution of the random variable

$$F = \frac{\frac{V}{m}}{\frac{U}{n}}.$$

We write $F \sim F_{m,n}$.

Note 16.1. Since the chi-square distribution is positive, so is the F distribution.

Definition 16.2. We define quantiles for the F distribution as follows: For $0 \leq \alpha \leq 1$, if $F \sim F_{m,n}$,

$$P(F \leq F_{\alpha,m,n}) = \alpha.$$

Obtaining the p.d.f. of an F random variable is a bit harder than for the t or χ^2 distributions, but it can be done as well:

Theorem 16.1. If $F \sim F_{m,n}$, then

$$f_F(w) = \frac{\Gamma\left(\frac{m+n}{2}\right) m^{m/2} n^{n/2} w^{m/2-1}}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)(n+mw)^{(n+m)/2}}$$

for $w \geq 0$, and $f_F(w) = 0$ otherwise.

Proof. The proof is involved, but here is the idea:

$$F_F(w) = P\left(\frac{V/m}{U/n} \leq w\right) = P\left(V \leq \frac{m}{n}Uw\right).$$

This is something we can easily compute:

$$P\left(V \leq \frac{m}{n}Uw\right) = \int_0^\infty \int_0^{\frac{m}{n}w} f_U(u) f_V(v)\, dv\, du.$$

From here on, it's just (nontrivial) algebra. □

16.2 Testing Equality of Variances

Recall that if $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ with μ and σ^2 unknown, then

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

So to test $H_0 : \sigma^2 = \sigma_0^2$ against, say, $H_1 : \sigma^2 > \sigma_0^2$ at the α significance level, we reject H_0 if $\frac{(n-1)s^2}{\sigma_0^2} \geq \chi^2_{(1-\alpha),n-1}$.

If dealing with $X_1, \ldots, X_n \sim N(\mu_X, \sigma_X^2), Y_1, \ldots, Y_m \sim N(\mu_Y, \sigma_Y^2)$, we know that

$$\frac{(n-1)S_X^2}{\sigma_X^2} \sim \chi^2_{n-1} \text{ and } \frac{(m-1)S_Y^2}{\sigma_Y^2} \sim \chi^2_{m-1},$$

so that

$$\frac{\frac{S_X^2}{\sigma_X^2}}{\frac{S_Y^2}{\sigma_Y^2}} \sim F_{n-1,m-1}.$$

In particular, under $H_0 : \sigma_X = \sigma_Y$,

$$\frac{S_X^2}{S_Y^2} \sim F_{n-1,m-1}.$$

Example 16.1. (Back to Random Dot Stereograms) Let's test, at the 5% significance level, for the equality of the variances of the two samples in the random dot stereogram example.

From the data $X_1, \ldots X_{43} \sim N(\mu_X, \sigma_X^2), Y_1, \ldots, Y_{35} \sim N(\mu_Y, \sigma_Y^2)$, we obtained $s_X^2 \approx 65.45$ and $s_Y^2 \approx 23.04$, which implies that

$$\frac{s_X^2}{s_Y^2} \approx 2.841.$$

A priori, there is no good reason for doing a 1-sided test rather than a 2-sided test, so we will test

$H_0 : \sigma_X = \sigma_Y$ against $H_1 : \sigma_X \neq \sigma_Y$.

Let's find the critical region for S_X^2/S_Y^2. We know that under $H_0, S_X^2/S_Y^2 \sim F_{42,34}$, so the critical region is

$$(0, F_{0.025,42,34}) \cup (F_{0.975,42,34}, \infty) \approx (0, 0.53) \cup (1.94, \infty).$$

Since

$$\frac{s_X^2}{s_Y^2} \approx 2.84 \in (0, 0.53) \cup (1.94, \infty),$$

we reject H_0. There is strong evidence that the variances of the two samples are not the same.

Math 4501 (Spring 2018) March 28, 2018
Prof. Christian Beneš

Lecture #17: Two-sample binomial test

17.1 Testing Equality of Binomial Parameters

We will assume that $X_1, \dots, X_n \sim Be(p_X)$ and $Y_1, \dots, Y_m \sim Be(p_Y)$ and $X = \sum_{i=1}^n X_i, Y = \sum_{j=1}^m Y_j$.

The key idea will be that

$$\frac{\frac{X}{n} - \frac{Y}{m} - (p_X - p_Y)}{\sqrt{\frac{p_X(1-p_X)}{n} + \frac{p_Y(1-p_Y)}{m}}} \overset{\text{approx}}{\sim} N(0,1).$$

In particular, if we test $H_0 : p_X = p_Y$, then under H_0 (using p for the common parameter),

$$\frac{\frac{X}{n} - \frac{Y}{m}}{\sqrt{\frac{p(1-p)}{n} + \frac{p(1-p)}{m}}} \overset{\text{approx}}{\sim} N(0,1).$$

This can't quite be used for hypothesis testing, as p is unknown, but we can approximate p by the estimator $\frac{X+Y}{n+m}$ which, under H_0 is unbiased, and write

$$\frac{\frac{X}{n} - \frac{Y}{m}}{\sqrt{\frac{\frac{X+Y}{n+m}(1-\frac{X+Y}{n+m})}{n} + \frac{\frac{X+Y}{n+m}(1-\frac{X+Y}{n+m})}{m}}} = \frac{\frac{X}{n} - \frac{Y}{m}}{\frac{1}{n+m}\sqrt{(X+Y)(n+m-(X+Y))\left(\frac{1}{n}+\frac{1}{m}\right)}} \overset{\text{approx}}{\sim} N(0,1).$$

Example 17.1. In a democracy, it is important to have as many people as possible registered to vote. A random sample of 1100 potential voters was placed into two groups:

- Group 1: 600 potential voters to whom registration reminders were sent; 332 registered.
- Group 2: 500 potential voters to whom no registration reminders were sent; 248 registered.

Determine at the 5% significance level whether reminders make a difference or not.

We let $X_1, \dots X_{600} \sim Be(p_X)$ be 1 if an individual of the first group registered, 0 otherwise and $Y_1, \dots Y_{500} \sim Be(p_Y)$ be 1 if an individual of the second group registered, 0 otherwise. We will test

$H_0 : p_X = p_Y$ against $H_1 : p_X > p_Y$.

Note that for this experiment, $x = 332, y = 248, n = 600, m = 500$.

The p-value equals

$$\begin{aligned} P & \left(\frac{X}{n} - \frac{Y}{m} > \frac{332}{600} - \frac{248}{500} \middle| H_0 \right) \\ &= P\left(\frac{\frac{X}{n} - \frac{Y}{m}}{\frac{1}{n+m}\sqrt{(X+Y)(n+m-(X+Y))\left(\frac{1}{n}+\frac{1}{m}\right)}} > \frac{\frac{332}{600} - \frac{248}{500}}{\frac{1}{n+m}\sqrt{(X+Y)(n+m-(X+Y))\left(\frac{1}{n}+\frac{1}{m}\right)}} \middle| H_0 \right) \\ &\approx P(Z \geq 1.8965) \approx 0.0287, \end{aligned}$$

so we reject H_0. There is strong evidence that registration reminders do increase the chance that a potential voter will register.

Math 4501 (Spring 2018) April 9 ,2018
Prof. Christian Beneš

Lecture #18: Confidence Intervals for the Two-Sample Problems

The procedure for finding confidence intervals in the two-sample case is based on the same distributional facts and approximations as in the construction of hypothesis tests.

To determine whether it is possible that

- $\mu_X = \mu_Y$
- $p_X = p_Y$
- $\sigma_X = \sigma_Y$

we will construct (approximate) confidence intervals for

- $\mu_X - \mu_Y$
- $p_X - p_Y$
- σ_X / σ_Y

Note 18.1. In the case of variances, the confidence interval is for the quotient, since that is the quantity which appears in the distributional fact involving the sample variances.

Example 18.1. (Random Dot Stereograms re-revisited) If $X_1, \dots, X_n \sim N(\mu_X, \sigma_X^2), Y_1, \dots, Y_m \sim N(\mu_Y, \sigma_Y^2)$ are independent, we know that

$$\frac{\frac{S_Y^2}{\sigma_Y^2}}{\frac{S_X^2}{\sigma_X^2}} \sim F_{m-1,n-1},$$

so

$$P\left(F_{\alpha/2,m-1,n-1} \leq \frac{\frac{S_Y^2}{\sigma_Y^2}}{\frac{S_X^2}{\sigma_X^2}} \leq F_{1-\alpha/2,m-1,n-1}\right) = 1 - \alpha.$$

Therefore,

$$P\left(F_{\alpha/2,m-1,n-1} \frac{S_X^2}{S_Y^2} \leq \frac{\sigma_X^2}{\sigma_Y^2} \leq F_{1-\alpha/2,m-1,n-1} \frac{S_X^2}{S_Y^2}\right) = 1 - \alpha$$

and so a $100(1-\alpha)\%$ confidence interval for $\frac{\sigma_X^2}{\sigma_Y^2}$ is

$$\left(F_{\alpha/2,m-1,n-1} \frac{S_X^2}{S_Y^2}, F_{1-\alpha/2,m-1,n-1} \frac{S_X^2}{S_Y^2}\right).$$

In our specific example, we had $S_X^2 \approx 65.45$ and $S_Y^2 \approx 23.04$ so that $S_X^2/S_Y^2 \approx 2.841$. Moreover, $F_{0.025,34,42} \approx 0.528, F_{0.975,34,42} \approx 1.94$, so a 95% confidence interval for σ_X^2/σ_Y^2 is (with $n = 43, m = 35$)

$$\big(F_{\alpha/2,m-1,n-1}2.841, F_{1-\alpha/2,m-1,n-1}2.841\big) \approx (1.5, 5.5).$$

Since 1 doesn't belong to this interval, we have good reason to suspect that $\sigma_X^2/\sigma_Y^2 \neq 1$, in other words, that $\sigma_X^2 \neq \sigma_Y^2$.

As a reminder of what this means (and doesn't mean), we can't say, based on our calculation, that there is a 95% chance that $\sigma_X^2/\sigma_Y^2 \neq 1$. This wouldn't make any sense since σ_X^2/σ_Y^2 isn't random so that we can't talk about probabilities. What we do know is that the random interval we constructed had a 95% chance of containing the true ratio σ_X^2/σ_Y^2. Since 1 doesn't belong to the interval, there are strong reasons for believing that $\sigma_X^2/\sigma_Y^2 \neq 1$ (though of course, we'll never know for sure).

The confidence intervals for $\mu_X - \mu_Y$ and $p_X - p_Y$ are constructed in a similar fashion. See the homework for more details.

Math 4501 (Spring 2018) April 16, 2018
Prof. Christian Beneš

Lecture #19: Goodness of Fit Tests

The general type of question addressed is: Could a given data set come from a given distribution?

A key concept here is the multinomial distribution:

Suppose $Y_1, \dots, Y_n$ are i.i.d. with $P(Y_i = r_j) = p_j, j = 1, \dots, t, \sum_{i=1}^t p_j = 1$. For $j = 1, \dots, t$, let

$$X_j = \sum_{i=1}^n \mathbb{1}\{Y_i = r_j\},$$

where

$$\mathbb{1}\{Y_i = r_j\} = \begin{cases} 1, & Y_i = r_j \\ 0, & \text{otherwise} \end{cases},$$

be the number of times r_j comes up. Then if $X_1 = x_1, \dots X_k = x_k$, we have $\sum_{i=1}^t x_j = n$.

Definition 19.1. The random vector $(X_1, \dots, X_t)$ has the **multinomial distribution** with parameters $n, p_1, \dots, p_t$.

Theorem 19.1. If $(X_1, \dots, X_t)$ has the multinomial distribution with parameters $n, p_1, \dots, p_t$ and $x_1, \dots, x_t \in \mathbb{N} \cup \{0\}$ with $\sum_{i=1}^t x_j = n$, then

$$P(X_1 = x_1, \dots X_k = x_k) = p_{(X_1, \dots, X_t)}(x_1, \dots, x_t) = \binom{n}{x_1, \dots, x_t} \prod_{i=1}^t p_i^{x_i}.$$

Proof. Each configuration of x_1 r_1s, $\dots, x_t$ r_ts has probability $\prod_{i=1}^t$. The number of such configurations is

$$\binom{n}{x_1}\binom{n-x_1}{x_2} \cdots \binom{n - \sum_{i=1}^{t-1}}{x_t} = \binom{n}{x_1, \dots, x_t}.$$

□

Theorem 19.2. If $(X_1, \dots, X_t)$ is a multinomial random variable with parameters $n, p_1, \dots, p_t$, then the marginal distribution of $X_j, j = 1, \dots, t$ is a binomial with parameters n and p_j.

Proof. X_j is the number of times r_j occurs. For each realization $Y_1, \dots, Y_n, P(r_j \text{ occurs}) = p_j$. Since $X_j = \sum_{i=1}^n \mathbb{1}\{Y_i = r_j\}$ is a sum of n Bernoulli random variables with parameter $p_j, X_j \sim Bin(n, p_j)$. □

Corollary 19.1. If $(X_1, \dots, X_t)$ is a multinomial random variable with parameters $n, p_1, \dots, p_t$, then

$$E[X_j] = np_j, \quad \text{Var}(X_j) = np_j(1 - p_j).$$

Example 19.1. (RDS re-re-revisited) We consider the times of the people who had visual and verbal aids. There were 35 samples with sample mean 5.55 and sample variance 23.04. Could it be that the data came from a normal distribution with mean 5.55 and variance 23.04?

The idea is to put the data in bins, say quartiles. For a standard normal, the quartiles are

$$-z_{0.25} \approx -0.674, z_{0.5} = -z_{0.5} = 0, z_{0.25} \approx 0.674.$$

Since if $X \sim N(5.55, 23.04)$, we can write

$$X = \sqrt{23.04}Z + 5.55,$$

and so the quartiles of X are

$$\sqrt{23.04}(-0.674) + 5.55 = 2.31, \quad \sqrt{23.04}(0) + 5.55 = 5.55, \quad \sqrt{23.04}(0.674) + 5.55 = 8.79.$$

So we can count the number of outcomes in each of the following four "bins":

$$(-\infty, 2.31], \quad (2.31, 5.55], \quad (5.55, 8.79], \quad (8.79, \infty).$$

We get, respectively, 10, 11, 8, and 6 realizations when, what we'd expect is $35/4 = 8.75$ realizations in each bin. How different is what we observed from that? How would we compute a p-value?

We will re-visit this example once we've taken another, less direct, approach to the question.

Math 4501 (Spring 2018) Monday, April 18, 2018
Prof. Christian Beneš

Lecture #20: Goodness of fit test

Following the motivation from our last class, we will develop tools to test whether a sample could come from a given distribution. The main tool is an ubiquitous object called *Pearson's test statistic.*

Theorem 20.1. Suppose $r_1, \dots, r_t$ are the possible outcomes (or ranges of outcomes) associated with n independent trials $Y_1, \dots, Y_n$,

- for $j = 1, \dots, n, P(Y_j = r_i) = p_i$ for $i = 1, \dots, t$,
- $X_i = \#\{\text{occurrences of } r_i\}, i = 1, \dots, t$.

Then

$$D = \sum_{i=1}^{t} \frac{(X_i - np_i)^2}{np_i} \overset{\text{approx.}}{\sim} \chi^2_{t-1}.$$

For a good approximation, we'd like $np_i \geq 5$ for all i and for a reasonable approximation, we'd like $np_i \geq 3/2$ for all i.

Proof. (for $t = 2$, that is, the 2-bin case) We want to show that

$$D = \sum_{i=1}^{t} \frac{(X_i - np_i)^2}{np_i} \approx Z^2$$

with $Z \sim N(0, 1)$. Note that

$$\begin{aligned} D &= \frac{(X_1 - np_1)^2}{np_1} + \frac{(n - X_1 - n(1 - p_1))^2}{n(1 - p_1)} \\ &= \frac{(X_1 - np_1)^2(1 - p_1) + (np_1 - X_1)^2 p_1}{np_1(1 - p_1)} \\ &= \frac{(X_1 - np_1)^2 \left((1 - p_1) + p_1\right)}{np_1(1 - p_1)} = \left(\frac{X_1 - np_1}{\sqrt{np_1(1 - p_1)}}\right)^2 \approx Z^2, \end{aligned}$$

by the central limit theorem. □

If for $1 \leq i \leq t, p_i$ is the probability of the outcome r_i and $p_{i,0}$ is some number, we now have a procedure for testing

$H_0 : p_1 = p_{1,0}, \dots, p_t = p_{t,0}$ against $H_1 : p_i \neq p_{i,0}$ for some $1 \leq i \leq t$:

If $x_1, \dots, x_t$ are the observed frequencies of the outcomes $r_1, \dots, r_t$ and $np_{1,0}, \dots np_{t,0}$ are the expected frequencies under H_0, then at the α level of significance, we should reject H_0 if

$$d = \sum_{i=1}^{t} \frac{(x_i - np_{i,0})^2}{np_{i,0}} \geq \chi^2_{1-\alpha, t-1}$$

and $np_{i,0} \geq 5$ for all i.

Note 20.1. We don't do a two-sided test here since if $d \leq \chi^2_{\alpha/2,t-1}$, then d is very close to 0, which means x_i and $np_{i,0}$ are close for all i, which would confirm H_0 rather than contradict it.

Example 20.1. A library wishes to ensure that its books fit the users' needs. It conducts and inventory and compares it to a sample of the books that are checked out. This gives

subject area	books in library(%)	number of books in sample
Business	32	268
Humanities	25	214
Natural Sciences	20	215
Social Sciences	15	115
Other	8	76

Use a 5% significance test to determine whether the distribution of checked out books fits the library's book distribution.

We assign a numerical value to each subject area and also include percentages of books in the sample to compare with the percentages of books in the library:

	subject area	books in library(%)	number of books in sample	books in sample(%)
1	Business	32	268	30.2
2	Humanities	25	214	24.1
3	Natural Sciences	20	215	24.2
4	Social Sciences	15	115	13
5	Other	8	76	8.5

We let p_i be the probability that a book from subject area i is picked in the sample and test

$H_0 : p_1 = p_{1,0} = 0.32, p_2 = p_{2,0} = 0.25, p_3 = p_{3,0} = 0.2, p_4 = p_{4,0} = 0.15, p_5 = p_{5,0} = 0.08$

against

$H_1 : p_i \neq p_{i,0}$ for some $1 \leq i \leq 5$.

Note first that $np_{i,0} \geq 5$ for all $1 \leq i \leq 5$.

Using the sample size $n = 888$, we compute

$$\begin{aligned} d &= \frac{(268-284.16)^2}{284.16} + \frac{(214-222)^2}{222} \frac{(215-177.6)^2}{177.6} \frac{(115-133.2)^2}{133.2} \frac{(76-71.04)^2}{71.04} \\ &\approx 11.92 > 9.488 = \chi^2_{0.95,4}, \end{aligned}$$

so we reject H_0. There is strong evidence suggesting that the books in the sample weren't selected according to the distribution of the books in the library.

Math 4501 (Spring 2018) April 23, 2018
Prof. Christian Beneš

Lecture #21: SECOND MIDTERM

Math 4501 (Spring 2018) Wednesday, April 25, 2018
Prof. Christian Beneš

Lecture #22: More Goodness of fit test

22.1 Goodness of Fit Tests with Unknown Parameters

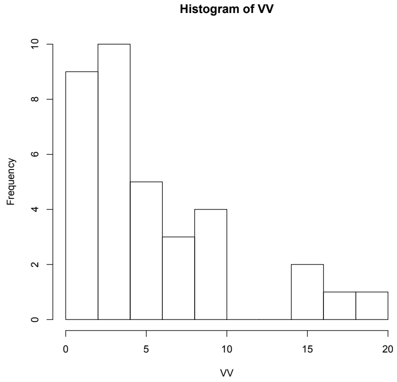
If we are given a data set, can we guess what distribution it might come from? Yes, to the extent that statistics allows us to make guesses. There are several methods that allow us to do this. We will look at *Pearson's goodness-of-fit test*.

Recall the Random-dot stereogram example in which we tried to determine if having information about the object hiding in a random dot stereogram increases the speed at which people recognize it. We performed a two-sample t-test which relied on equality of variances and on the normality of the data. At the time we didn't know how to test for either of these hypotheses. We already saw that the there was cause to reject equality of variances, using the F test. Had we not rejected equality of variances, we would still have needed to check normality of the data. We will see in a few moments how to do this, but let's first try to get some intuition for what the answer might be by looking at the histogram obtained from the data. The times needed for people with visual aid to recognize the image can be found at

http://userhome.brooklyn.cuny.edu/cbenes/RDS.txt

We can import the file into R and draw the histogram with the following sequence of commands:

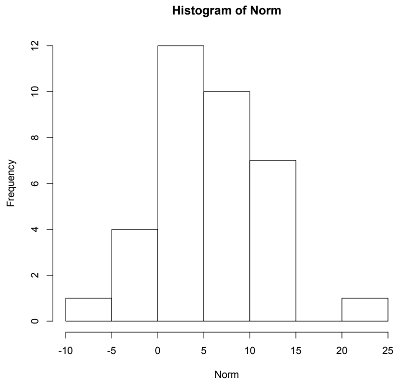
```
> RDS=“http://userhome.brooklyn.cuny.edu/cbenes/RDS.txt”
> Visual=scan(RDS)
> hist(Visual)
```



The question is now: Could this be the the histogram of 35 independent realizations of a normal r.v.?

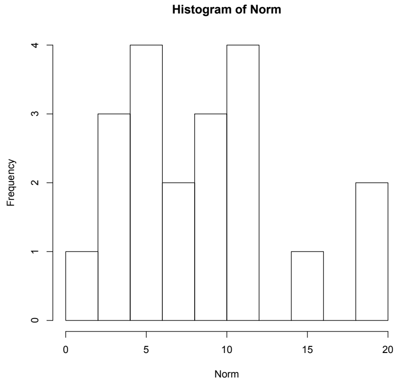
And what about the following graph?

```
> Norm=rnorm(35,5.55,4.80)
> hist(Norm)
```



And the following?

```
> Norm=rnorm(20,5.55,4.80)
> hist(Norm)
```



It turns out that the last two graphs are of independent normals (35 and 20, respectively). With small data sets, one should be careful not to draw conclusions too quickly, as even perfectly normal data sets may not appear to be so.

Let us now develop the tools that will allow us to answer this question precisely.

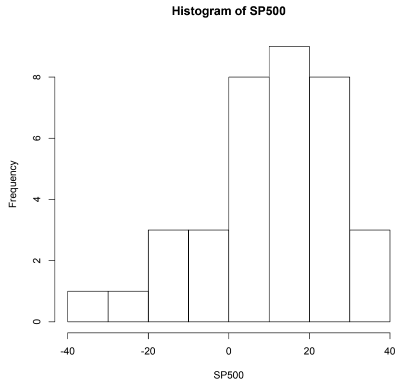
Let's look at another data set, that of the Percent Gain or Loss for the S&P 500 Index Yearly Returns for the years 1975 to 2010, available at

http://userhome.brooklyn.cuny.edu/cbenes/S&P500Returns.txt

(Calculations do not reflect any dividends paid or any stock spinoffs from original stock. Taxes and commissions are not factored into calculations.)

As before, we can acquire the data set as follows and plot a histogram:

```
> SP="http://userhome.brooklyn.cuny.edu/cbenes/S&P500Returns.txt"
> SP500=scan(SP)
> hist(SP500)
```



We can also obtain the sample mean and sample standard deviation for the data using the following commands.

```
> mean(SP500)
[1] 9.795
> sd(SP500)
[1] 16.64338
```

Now if this histogram were that of 36 realizations of normal random variables, what would be our best guess as to the mean and standard deviation of those normals? The same as the sample mean and sample standard deviation, of course!

So now the question is: What would we expect 36 samples from a $N(9.8, 277)$ to look like? Well, we'd certainly like roughly half to be on each side of 9.8, but thanks to normal quantiles, we can say much more about what we'd expect. Let's look at how we'd do this by focusing on quartiles:

We know from the normal table that $z_{0.25} = 0.6745$. Therefore, a standard normal has probability 25% of falling in each of the following intervals:

$$(-\infty, -0.6745) \quad (-0.6745, 0) \quad (0, 0.6745) \quad (0.6745, \infty).$$

So using the fact that if $X \sim N(9.8, 277), X = 16.64338Z + \mu$, we see that X has probability 25% of falling in each of the following intervals:

$$(-\infty, -0.6745 \cdot 16.643 + 9.8) \quad (-0.6745 \cdot 16.643 + 9.8, 0 \cdot 16.643 + 9.8)$$

$$(0 \cdot 16.643 + 9.8, 0.6745 \cdot 16.643 + 9.8) \quad (0.6745 \cdot 16.643 + 9.8, \infty),$$

equivalently of falling in each of the following intervals:

$$(-\infty, -1.43) \quad (-1.43, 9.8) \quad (9.8, 21.03) \quad (21.03, \infty).$$

So if the Percent Gain or Loss for the S&P 500 Index Yearly Returns for the years 1975 to 2010 were normally distributed, we would expect to find about 9 data points in each of the intervals ("bins") above. It turns out that there are 8, 8, 10, and 10, respectively. Is this evidence against the hypothesis of normality?

A natural question to ask would be: What is the probability that we would have observed this? If for $i = 1, \ldots, 4, X_i$ is the number of realizations in bin i, what is

$$P(X_1 = 8, X_2 = 8, X_3 = 10, X_4 = 10)?$$

We can compute this exactly, since (X_1, X_2, X_3, X_4) is multinomial with parameters $36, 1/4$, $1/4, 1/4, 1/4$.

$$P(X_1 = 8, X_2 = 8, X_3 = 10, X_4 = 10) = \binom{36}{8, 8, 10, 10} \left(\frac{1}{4}\right)^3 6.$$

More importantly, what is the probability that *something as extreme or more as what we observed* would have happened?

It's not clear what that means... so we'll try a different approach. The key idea is to construct a test statistic from the random variables $X_1, \ldots, X_t$ in such a way that its distribution can be found or at least approximated.

Theorem. (Pearson's test statistic) If $r_1, \ldots, r_t$ are the possible (ranges of) outcomes associated with n independent trials from an entirely known distribution, $P(r_i) = p_i$, and $X_i = \#(\text{outcomes of } r_i), i = 1, \ldots, t$, then

$$D = \sum_{i=1}^{t} \frac{(X_i - np_i)^2}{np_i} \overset{\text{approx}}{\sim} \chi^2_{t-1}.$$

Note that we don't generally know the parameters of the distribution which we would like to fit to the data. When we have to estimate them, the statement of the theorem changes just a little bit:

Theorem. (Pearson's test statistic) If $r_1, \ldots, r_t$ are the possible (ranges of) outcomes associated with n independent trials from a distribution with s unknown parameters estimated with the maximum likelihood estimator, $\hat{p}_i$ is the estimated probability of r_i, and $X_i = \#(\text{outcomes of } r_i), i = 1, \ldots, t$, then

$$D = \sum_{i=1}^{t} \frac{(X_i - n\hat{p}_i)^2}{n\hat{p}_i} \overset{\text{approx}}{\sim} \chi^2_{t-1}.$$

Example: For the S&P500 returns, let's test at the 95% level if the data could be normal with estimated mean and variance. Note that the last theorem requires using the maximum likelihood estimates for the mean and variance (under the normality assumption). While the sample mean is the maximum likelihood estimator for the mean, the sample variance is not the maximum likelihood estimator for the variance, so we need a slightly different estimate from the one above. We need to use the maximum likelihood estimator for the variance which, in R, is

```
> s2=sum((SP500-mean(SP500))2)/36
```

so that the variance is

```
> s=sqrt(sum((SP500-mean(SP500))2)/36)
> s
[1] 16.41059
```

We can then find our bins, as above, using the commands for normal quantiles as follows:

```
> s*qnorm(0.25)+mean(SP500)
[1] -1.273777
> s*qnorm(0.75)+mean(SP500)
[1] 20.86378
```

This means that our four equiprobable bins for the normal distribution with estimated parameters are

$$(-\infty, -1.273777), \quad (-1.273777, 9.795), \quad (9.795, 20.86378), \quad (20.86378, \infty).$$

We have $n = 36$ and need $np_i \geq 5$ for all i. Though this is not optimal, let's choose $p_i = 4$ as above. Then

$$d_1 = \frac{1}{9} + \frac{1}{9} + \frac{1}{9} + \frac{1}{9} = \frac{4}{9} \not\geq \chi^2_{0.95,1} = 3.841.$$

Therefore, there is no cause to reject the hypothesis that the data are normally distributed with estimated mean and variance.

We could improve our analysis by taking the number of bins to be as large as possible. By making the p_i equal, we can choose $p_i = 1/7$, so that for every i, $np_i = \frac{36}{7} \geq 5$. Then, using the commands

```
> s*qnorm(1/7)+mean(SP500)
[1] -7.724465
> s*qnorm(2/7)+mean(SP500)
[1] 0.5074442
> s*qnorm(3/7)+mean(SP500)
[1] 6.84089
> s*qnorm(4/7)+mean(SP500)
[1] 12.74911
> s*qnorm(5/7)+mean(SP500)
[1] 19.08256
> s*qnorm(6/7)+mean(SP500)
[1] 27.31447
```

we see that our bins are

$$(\infty, -7.724), (-7.724, 0.507), (0.507, 6.841), (6.841, 12.749),$$

$$(12.749, 19.083), (19.083, 27.314), (27.314, \infty)$$

For $i = 1, \ldots, 7$, we call these intervals I_i and r.v. counting the number of data points in each X_i. Then

$$x_1 = 6, x_2 = 2, x_3 = 6, x_4 = 4, x_5 = 5, x_6 = 10, x_7 = 3.$$

Therefore,

$$d_2 = \sum_{i=1}^{7} \frac{(x_i - 36/7)^2}{36/7} \approx 7.94 \not> 9.844 = \chi_{0.05,4}.$$

Therefore, there is no cause to reject the hypothesis that the data are normally distributed with estimated mean and variance.

Math 4501 (Spring 2018) Monday, April 30, 2018
Prof. Christian Beneš

Lecture #23: Least Squares Estimation

24 Contingency Tables

We will see in this section how one can test for independence of two variables.

Example 24.1. Consider the following table:

Happiness Level Income Level	low	moderate	high
below average	83	249	94
average	221	372	53
above average	110	159	21

We will try to use this table to determine if the variables "Income level" and "Happiness level" may be independent or not. We re-write the table with totals and numerical values assigned to the various outcomes of the two variables:

Happiness Level Income Level	low (1)	moderate (2)	high (3)	
below average (1)	83	249	94	426
average (2)	221	372	53	646
above average (3)	110	159	21	290
	414	780	168	1362

Can we infer from this table whether the (random) variables X=income level and Y=happiness level are independent or not?

There are two key ideas involved:

Key idea 1: If X and Y are independent, then for all i, j,

$$P(X = i, Y = j) = P(X = i)P(Y = j).$$

Note that from here on $\hat{P}$ represents estimated probabilities and $\hat{E}$ estimated expectations.

For our sample we find the estimates

$$\begin{aligned}
\hat{P}(X = 1) &= \frac{426}{1362} \approx 0.313, \quad \hat{P}(Y = 1) = \tfrac{414}{1362} \approx 0.304,\\
\hat{P}(X = 2) &= \frac{646}{1362} \approx 0.474, \quad \hat{P}(Y = 2) = \tfrac{780}{1362} \approx 0.573,\\
\hat{P}(X = 3) &= \frac{290}{1362} \approx 0.213, \quad \hat{P}(Y = 3) = \tfrac{168}{1362} \approx 0.123.
\end{aligned} \tag{5}$$

In particular, if we pick a person at random, we are drawing from the estimated joint distribution defined by (5) and the assumed independence. Then if drawing n samples $(X,Y)_\ell, 1 \le \ell \le n$, we get

$$\begin{aligned}
E[\#(\text{times } (X,Y)_\ell = (i,j))] &= E[\sum_{\ell=1}^{n} \mathbb{1}\{(X,Y)_\ell = (i,j)\}] \\
&= \sum_{\ell=1}^{n} E[\mathbb{1}\{(X,Y)_\ell = (i,j)\}] \\
&= \sum_{\ell=1}^{n} P((X,Y)_\ell = (i,j)) \\
&\overset{(X,Y)_\ell \text{i.d.}}{=} nP((X,Y)_\ell = (i,j)) \\
&\overset{X,Y \text{ indep.}}{=} nP(X=i)P(Y=j).
\end{aligned}$$

<u>Key idea 2:</u> If n observations are taken on a sample space partitioned by $A_1, \dots, A_r$ and also by $B_1, \dots B_c$. For $1 \le i \le r, 1 \le j \le c$, let

$$p_i = P(A_i), q_j = P(B_j), p_{ij} = P(A_i B_j).$$

If X_{ij} is the number of observations of $\{X \in A_i, Y \in B_j\}$, then

$$D_2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(X_{ij} - np_{ij})^2}{np_{ij}} \overset{\text{approx.}}{\sim} \chi^2_{rc} \quad (\text{if } np_{ij} \ge 5 \text{ for all } i,j).$$

Also, if $H_0 : A_i$ are independent of B_j and $\hat{P}_i$ and $\hat{Q}_j$ are the estimated probabilities of A_i and B_j, respectively, then

$$D_2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(X_{ij} - n\hat{P}_i\hat{Q}_j)^2}{n\hat{P}_i\hat{Q}_j} \overset{\text{approx.}}{\sim} \chi^2_{(r-1)(c-1)} \quad (\text{if } np_{ij} \ge 5 \text{ for all } i,j).$$

In our example, since we have

$$\begin{aligned}
n\hat{P}(X=1)\hat{P}(Y=1) &= 1362 \frac{426}{1362} \frac{414}{1362} \approx 129.5, \\
n\hat{P}(X=1)\hat{P}(Y=2) &= 1362 \frac{426}{1362} \frac{780}{1362} \approx 244, \\
n\hat{P}(X=1)\hat{P}(Y=3) &= 1362 \frac{426}{1362} \frac{168}{1362} \approx 52.2, \\
n\hat{P}(X=2)\hat{P}(Y=1) &= 1362 \frac{646}{1362} \frac{414}{1362} \approx 196.4, \\
n\hat{P}(X=2)\hat{P}(Y=2) &= 1362 \frac{646}{1362} \frac{780}{1362} \approx 370, \\
n\hat{P}(X=2)\hat{P}(Y=3) &= 1362 \frac{646}{1362} \frac{168}{1362} \approx 79.7
\end{aligned}$$

$$n\hat{P}(X=3)\hat{P}(Y=1) = 1362\frac{290}{1362}\frac{414}{1362} \approx 88.1,$$
$$n\hat{P}(X=3)\hat{P}(Y=2) = 1362\frac{290}{1362}\frac{780}{1362} \approx 166.1,$$
$$n\hat{P}(X=3)\hat{P}(Y=3) = 1362\frac{290}{1362}\frac{168}{1362} \approx 35.8.$$

$$\begin{aligned} d_2 &= \frac{(83-129.5)^2}{129.5} + \frac{(249-244)^2}{244} + \frac{(94-52.2)^2}{52.2} \\ &+ \frac{(221-196.4)^2}{196.4} + \frac{(372-370)^2}{370} + \frac{(53-79.7)^2}{79.7} \\ &+ \frac{(110-88.1)^2}{88.1} + \frac{(159-166.1)^2}{166.1} + \frac{(21-35.8)^2}{35.8} \\ &\approx 16.7 + 8 \text{ terms} > 9.488 \approx \chi^2_{0.95,4}, \end{aligned}$$

so we reject H_0 at the 5% level. There is strong evidence that income levels and happiness levels are correlated.

24.1 Least squares linear regression

Consumer reports picked 8 cars at random and produced the following table:

weight (1000s of lbs)	miles per gallon
x	y
27	30
44	19
32	24
47	13
23	29
40	17
34	21
52	14

If we plot x against y, can we find a straight line that fits the data well?

Q: What does it mean to "fit the data well"?

A: There are many possible answers:

- A line that contains as many of the data points as possible (more than two such points is typically not possible)
- A line such that the sum of the distances between the data points and the line is minimal.

- A line such that the sum of squares of the vertical distances is minimal (sounds far fetched, but...)
- ...

We will focus on the last of these (and will see in a few lectures why this is a natural choice).

Our goal will be to find a line $y = a + bx$ (i.e., find a and b) such that for the n points $(x_1, y_1), \dots (x_n, y_n)$, the quantity

$$L(a,b) = \sum_{i=1}^{n} (y_i - (a + bx_i))^2$$

is minimal.

Theorem 24.1. Given n points $(x_1, y_1), \dots (x_n, y_n)$, the straight line $y = a + bx$ which minimizes $L = \sum_{i=1}^{n}(y_i - (a + bx_i))^2$ has slope

$$b = \frac{n \sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2}$$

and y-intercept

$$a = \frac{\sum_{i=1}^{n} y_i - b \sum_{i=1}^{n} x_i}{n} = \bar{y} - b\bar{x}.$$

Proof. $L(a,b) = \sum_{i=1}^{n}(y_i - (a + bx_i))^2$ is quadratic in a and b, so we can find its maximum by solving

$$\frac{\partial L}{\partial b}(a,b) = 0, \quad \frac{\partial L}{\partial a}(a,b) = 0.$$

$$\frac{\partial L}{\partial b}(a,b) = 0 \iff \sum_{i=1}^{n} -2x_i(y_i - (a + bx_i)) = 0 \iff \sum_{i=1}^{n} x_i(y_i - (a + bx_i)) = 0$$

and

$$\frac{\partial L}{\partial a}(a,b) = 0 \iff \sum_{i=1}^{n} -2(y_i - (a + bx_i)) = 0 \iff \sum_{i=1}^{n} (y_i - (a + bx_i)) = 0.$$

This yields the following two linear equations in a and b:

$$\sum_{i=1}^{n} x_i y_i = a \sum_{i=1}^{n} x_i + b \sum_{i=1}^{n} x_i^2, \quad \sum_{i=1}^{n} y_i = na + b \sum_{i=1}^{n} x_i.$$

Solving for a and b (for instance by substitution or Cramér's law from linear algebra) yields the theorem □

In the example given at the beginning of this lecture, we find $a \approx 43.33$ and $b \approx -0.60$ as we can obtain from R using the following commands:

```
> X=c(27,44,32,47,23,40,34,52)
> Y=c(30,19,24,13,29,17,21,14)
> fit=lm(Y∼X)
> fit
Call:
lm(formula = Y ∼ X)
Coefficients:
(Intercept) X
43.3263 -0.6007
```

We can see how well the line fits by using the following commands:

```
> plot(X,Y)
> abline(fit)
```

24.2 Least squares for power and exponential functions

How do we find the functions which best approximates the data $(x_i, y_i), i = 1, \ldots, n$, when

1. $f(x) = ae^{bx}$,
2. $f(x) = ax^b$?

1. The key idea is that if $y = f(x) = ae^{bx}$, then $\ln(y) = \ln(a) + bx$, so $\ln(y)$ is a linear function of x. So using the least squares result for linear functions, we get

$$b = \frac{n\sum_{i=1}^n x_i \ln(y_i) - (\sum_{i=1}^n x_i)(\sum_{i=1}^n \ln(y_i))}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2},$$

$$\ln(a) = \overline{\ln(y)} - b\bar{x}.$$

2. The key idea is that if $y = f(x) = ax^b$, then $\ln(y) = \ln(a) + b\ln(x)$, so $\ln(y)$ is a linear function of $\ln(x)$. So using the least squares result for linear functions, we get

$$b = \frac{n\sum_{i=1}^n \ln(x_i)\ln(y_i) - (\sum_{i=1}^n \ln(x_i))(\sum_{i=1}^n \ln(y_i))}{n\sum_{i=1}^n (\ln(x_i))^2 - (\sum_{i=1}^n \ln(x_i))^2},$$

$$\ln(a) = \overline{\ln(y)} - b\overline{\ln(x)}.$$

Example 24.2. (log plots and log-log plots)

```
> A=1:20
> B=A^2/10
> C=1.1^A
> plot(A,B)
> plot(A,C)
> plot(A,log(B))
> plot(A,log(C))
> plot(log(A),log(B))
> plot(log(A),log(C))
```

Math 4501 (Spring 2018) Wednesday, May 3, 2017
Prof. Christian Beneš

Lecture #24: Least Squares Estimation; the Kolmogorov-Smirnov test

24.1 Residuals

Definition 24.1. The *residuals* in a least square model are the difference between the data and the corresponding points on the least square line. More precisely, if the data are $(x_i, y_i), i = 1, \ldots, n$ and the least squares line is $y = a + bx$, then the residuals are

$$r_i = y_i - (a + bx_i).$$

A graph of x_i vs. r_i is a *residual plot.*

A residual plot can indicate whether the assumption about the type of curve assumed to be underlying the data was well chosen. For a good choice, the residual plot should look completely random.

Example 24.1. In our example from earlier, we can obtain the residuals from R as follows:

```
> X=c(27,44,32,47,23,40,34,52)
> Y=c(30,19,24,13,29,17,21,14)
> fit=lm(Y~X)
> fit$resid
> plot (X,fit$resid)
```

A quick inspection of the residuals shows no obvious pattern, so it is reasonable to assume that the least squares line is an adequate model for the data.

Example 24.2. A perfectly alternating residual plot would also exhibit a pattern which should cause us to be cautious with the model.

Example 24.3. Consider the power function data from our last class. We will try to find a least squares line for it:

```
> A=1:20
> B=A^2/10
> powerfit=lm(B~A)
> plot(A,powerfit$resid)
```

24.2 The Kolmogorov-Smirnov Test

We already know how to test if a sample could have been produced by a given distribution using Pearson's chi-square goodness-of-fit test. We will now see another method which allows us to test the same thing.

We will assume that $x_1, \ldots, x_n$ are a sample from a continuous distribution, so that $x_{(1)} < x_{(2)} < \cdots x_{(n)}$.

Definition 24.2. The *empirical distribution* of the sample $X_1, \ldots, X_n$ is

$$F_n(x) = \begin{cases} 0, & x < x_{(1)} \\ k/n, & x_{(k)} \leq x < x_{(k+1)} \\ 1, & x_{(n)} \leq x \end{cases} .$$

Note that the empirical function depends on the data, so is a priori random. How can we use this object to determine if a given data set could be generated by a specific distribution?

Fix x and let $W = F_n(x)$. Then W is a random variable which can take values $0, 1/n, 2/n, \ldots, 1$.

Note that $nW = k \iff k$ observations are $\leq x$ and $n-k$ observations are $> x$.

The probability that a single observation is $\leq x$ is $F(x)$, so by independence of trials,

$$P(nW = k) = \binom{n}{k} F(x)^k (1 - F(x))^{n-k}, k = 0, \ldots, n.$$

Note that since $nW \sim Bin(n, F(x))$,

$$E[nW] = nF(x), \quad \operatorname{Var}(nW) = nF(x)(1 - F(x)).$$

Therefore,

$$E[F_n(x)] = E[W] = F(x), \quad \operatorname{Var}(F_n(x)) = \operatorname{Var}(W) = \frac{F(x)(1 - F(x))}{n}.$$

This suggests that for every $x, F_n(x) \to F(x)$.

Definition 24.3. The Kolmogorov-Smirnov statistic is

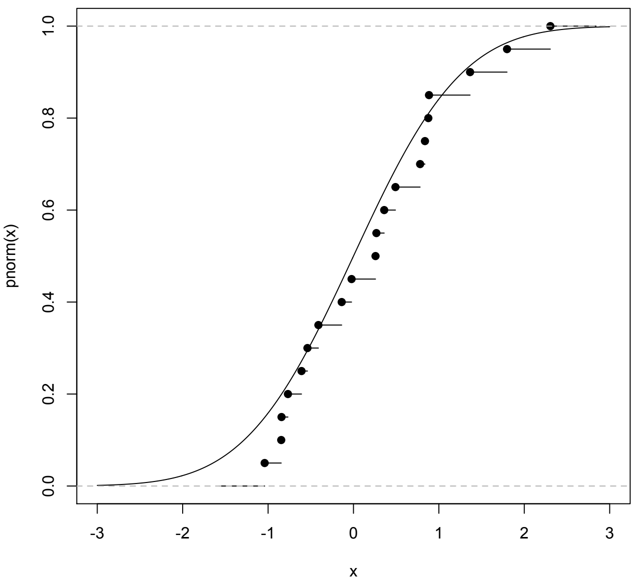
$$D_n = \sup_x \left[|F_n(x) - F_0(x)| \right].$$

Its distribution can be computed (R does it for you), which is all one needs to test hypotheses.

We now work out an example in R:

Example 24.4. We will generate 20 standard normal samples and compare the empirical cumulative distribution with the standard normal cdf. We also use the Komogorov-Smirnov test to check if the data could come from a standard normal distribution (presumably, the test should not lead to the rejection of that hypothesis, since we do know that the data are standard normal).

```
> x=seq(-3,3,0.01)
> plot(x,pnorm(x),type=”l”)
> N=rnorm(20)
> ecdf(N)
> plot(ecdf(N),add=TRUE)
```



```
> ks.test(N,pnorm)
One-sample Kolmogorov-Smirnov test
data: N
D = 0.1518, p-value = 0.6907
alternative hypothesis: two-sided
```

Since the p-value is greater than 5%, we fail to reject

H_0: The data follow a standard normal distribution.

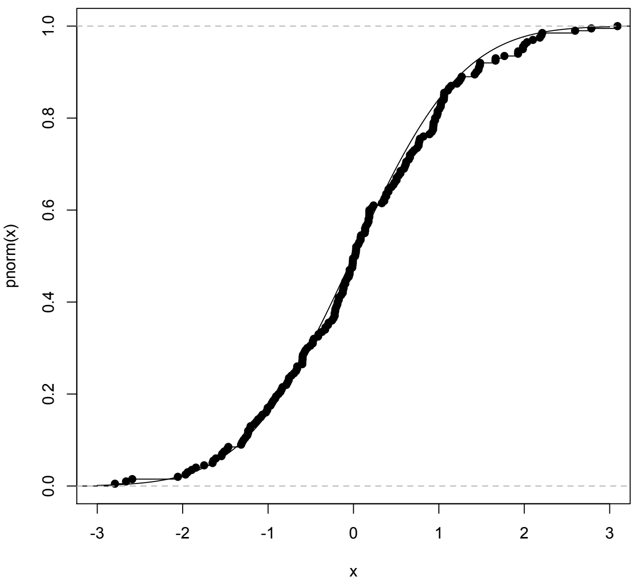
against

H_1: They don't.

at the 5% significance level.

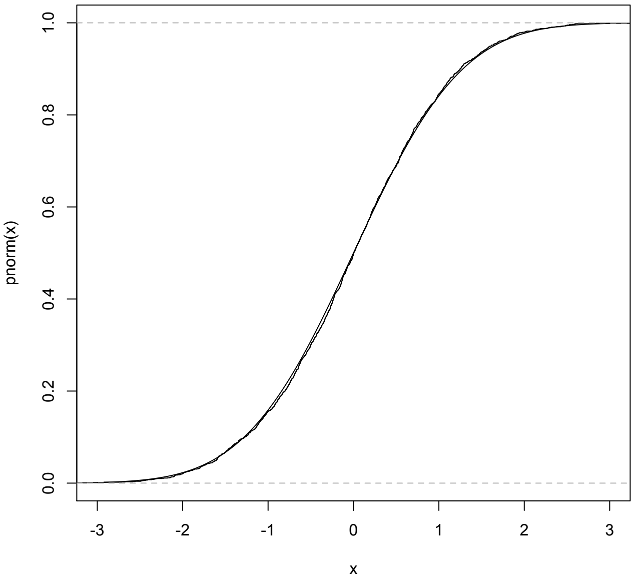
Note that the empirical distribution function should look more and more like the true distribution function as the number of samples increases. Here is a picture of the cdf and ecdf for 200 standard normal samples:

```
> x=seq(-3,3,0.01)
> plot(x,pnorm(x),type="l")
> N=rnorm(200)
> ecdf(N)
> plot(ecdf(N),add=TRUE)
```



And another for 2000 standard normal samples:

```
> x=seq(-3,3,0.01)
> plot(x,pnorm(x),type=”l”)
> N=rnorm(2000)
> ecdf(N)
> plot(ecdf(N),add=TRUE)
```



Math 4501 (Spring 2018) Monday, May 8, 2017
Prof. Christian Beneš

Lecture #25: The Linear Model

25.1 The Linear Model

The least squares model doesn't take into account any randomness as the variables x and y were not given any randomness structure. In the linear model, we will think of our data as a combination of a deterministic and a random component.

Example 25.1. If a sample is taken from a population and x is the number of years of education is plotted against y, annual income, then it is not unreasonable to assume that for fixed x, the average value of y is given $y = 4000x - 20000$.

We will now think of y as being a random quantity that depends on the deterministic value x, so we'll write Y rather than y. For each x, we have a conditional distribution $f_{Y|x}$. Then $E[Y|x]$, the expectation associated with the density $f_{Y|x}$ is the *regression curve* on x (also called the *trend*).

The *simple linear model* satisfies

1. $Y|x \sim N(\mu_x, \sigma^2)$ (with the same σ for all x).
2. $y = E[Y|x] = \beta_0 + \beta_1 x$.
3. $Y|x$ is independent of $Y|x'$ if $x \neq x'$.

Show a graphical representation for the example above.

Our goal will be to estimate β_0, β_1, and σ^2 in the simple linear model.

25.2 Maximum Likelihood Estimators for β_0, β_1, and σ^2

By thinking of $Y_1, \dots, Y_n$ as random variables, we get points $(x_1, Y_1), \dots (x_n, Y_n)$ of which we assume that they come from the simple linear model with $E[Y|x] = \beta_0 + \beta_1 x$. We can then find estimators for β_0, β_1, and σ^2 in terms of the random variables $Y_1, \dots, Y_n$ (and estimates in terms of the outcomes $y_1, \dots, y_n$).

Theorem 25.1. In the simple linear model, the Maximum Likelihood Estimators for $\beta_0, \beta_1, \sigma^2$ are

$$\hat\beta_1 = \frac{n\sum_{i=1}^n x_i Y_i - (\sum_{i=1}^n x_i)(\sum_{i=1}^n Y_i)}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}$$

$$\hat\beta_0 = \bar Y - \hat\beta_1 \bar x$$

$$\hat\sigma^2 = \frac{1}{n}\sum_{i=1}^n (Y_i - \hat Y_i)^2, \text{ where } \hat Y_i = \hat\beta_0 + \hat\beta_1 x_i$$

Note 25.1.

1. In particular, given the data $(x_1, y_1), \ldots, (x_n, y_n)$, the Maximum Likelihood Estimates $\beta_{0,e}, \beta_{1,e}$ for β_0 and β_1 are the same as the least squares estimates.

2. Why is the linear model better than just considering least squares estimates? Since $\hat{\beta}_0$ and $\hat{\beta}_1$ are random variables, we can use them to find confidence intervals or to test hypotheses for β_0 and β_1.

Proof. To find maximum likelihood estimators, we of course first need the likelihood function.

$$L(y_1, \ldots, y_n) = \prod_{i=1}^{n} f_{Y|x_i}(y_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{y_i - \beta_0 - \beta_1 x_i}{\sigma}\right)^2}.$$

Therefore,

$$\ln(L(y_1, \ldots, y_n)) = -\frac{1}{2}\sum_{i=1}^{n}\left(\frac{y_i - \beta_0 - \beta_1 x_i}{\sigma}\right)^2 - \frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2).$$

Therefore,

$$\frac{\partial}{\partial \beta_0}(\ln(L)) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial}{\partial \beta_1}(\ln(L))\frac{1}{\sigma^2}\sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial}{\partial \sigma^2}(\ln(L)) = \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 - \frac{n}{2\sigma^2} = 0$$

Now checking that $\beta_{0,e}, \beta_{1,e}$, and σ_e^2 solve this equation completes the proof. □

Note 25.2. In the textbook, $\hat{\beta}_0, \hat{\beta}_1$, and $\hat{\sigma^2}$ are estimators when boldfaced, estimates otherwise.

To test hypotheses for β_0, β_1, and σ^2 we need to know the distributions of estimators for these parameters. The next theorem determines the exact distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$.

Theorem 25.2.

(a) $\hat{\beta}_0$ and $\hat{\beta}_1$ are normally distributed.

(b) $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased.

(c)

$$\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

(d)

$$\mathrm{Var}(\hat{\beta}_0) = \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Proof. (for $\hat{\beta}_1$)

$$\begin{aligned}
\hat{\beta}_1 &= \frac{n\sum_{i=1}^n x_iY_i - (\sum_{i=1}^n x_i)(\sum_{i=1}^n Y_i)}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} = \frac{n\sum_{i=1}^n x_iY_i - n^2\bar{x}\bar{Y}}{n\sum_{i=1}^n x_i^2 - n^2\bar{x}^2} \\
&= \frac{\sum_{i=1}^n x_iY_i - n\bar{x}\bar{Y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} \overset{\sum_{i=1}^n 2x_i\bar{x}=2n\bar{x}^2}{=} \frac{\sum_{i=1}^n x_iY_i - \bar{x}\sum_{i=1}^n Y_i - \bar{Y}\sum_{i=1}^n x_i + n\bar{x}\bar{Y}}{\sum_{i=1}^n(x_i^2 - 2x_i\bar{x} + \bar{x}^2)} \\
&= \frac{\sum_{i=1}^n(x_i-\bar{x})Y_i - \bar{Y}\sum_{i=1}^n(x_i-\bar{x})}{\sum_{i=1}^n(x_i-\bar{x})^2} \overset{\sum_{i=1}^n(x_i-\bar{x})=0}{=} \frac{\sum_{i=1}^n(x_i-\bar{x})Y_i}{\sum_{i=1}^n(x_i-\bar{x})^2} = \sum_{i=1}^n \frac{(x_i-\bar{x})}{\sum_{i=1}^n(x_i-\bar{x})^2}Y_i.
\end{aligned}$$

Since the summands are independent normals, we see that $\hat{\beta}_1$ is normal, since it is the sum of independent normals. Moreover, our calculations yield

$$\begin{aligned}
E[\hat{\beta}_1] &= \sum_{i=1}^n \frac{(x_i-\bar{x})}{\sum_{i=1}^n(x_i-\bar{x})^2}E[Y_i] = \sum_{i=1}^n \frac{(x_i-\bar{x})}{\sum_{i=1}^n(x_i-\bar{x})^2}(\beta_0+\beta_1x_i) \\
&= \beta_0\sum_{i=1}^n \frac{(x_i-\bar{x})}{\sum_{i=1}^n(x_i-\bar{x})^2} + \sum_{i=1}^n \frac{\beta_1x_i(x_i-\bar{x})}{\sum_{i=1}^n(x_i-\bar{x})^2} \\
&\overset{\sum_{i=1}^n(x_i-\bar{x})=0, \sum_{i=1}^n(x_i-\bar{x})\beta_1\bar{x}=0}{=} \frac{1}{\sum_{i=1}^n(x_i-\bar{x})^2}\left(\sum_{i=1}^n(x_i-\bar{x})\beta_1x_i - \sum_{i=1}^n(x_i-\bar{x})\beta_1\bar{x}\right) \\
&= \frac{1}{\sum_{i=1}^n(x_i-\bar{x})^2}\sum_{i=1}^n(x_i-\bar{x})\beta_1(x_i-\bar{x}) = \beta_1.
\end{aligned}$$

Also,

$$\begin{aligned}
\mathrm{Var}(\hat{\beta}_1) &= \mathrm{Var}(\sum_{i=1}^n \frac{(x_i-\bar{x})}{\sum_{i=1}^n(x_i-\bar{x})^2}Y_i) \\
&= \frac{1}{(\sum_{i=1}^n(x_i-\bar{x})^2)^2}\sum_{i=1}^n(x_i-\bar{x})^2\,\mathrm{Var}(Y_i) = \frac{\sigma^2}{\sum_{i=1}^n(x_i-\bar{x})^2}
\end{aligned}$$

□

Math 4501 (Spring 2018) Wednesday, May 10, 2017
Prof. Christian Beneš

Lecture #26: The Linear Model

26.1 Inference for σ^2

Theorem 26.1. (a) $\hat{\beta}_1, \bar{Y}$, and $\hat{\sigma}^2$ are mutually independent.

(b) $\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}$.

Note 26.1. Part (b) of the last theorem is all you need to make inference for σ^2.

Corollary 26.1. $S^2 = \frac{n}{n-2}\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$ is an unbiased estimator for σ^2.

Proof.

$$E\left[\frac{n}{n-2}\hat{\sigma}^2\right] = \frac{n}{n-2}E\left[\hat{\sigma}^2\right] = \frac{n}{n-2}\frac{\sigma^2}{n}E\left[\frac{n\hat{\sigma}^2}{\sigma^2}\right] \overset{E\left[\frac{n\hat{\sigma}^2}{\sigma^2}\right]=n-2}{=} \sigma^2.$$

□

26.2 Inference for β_1

Theorem 26.2.

$$\frac{\hat{\beta}_1 - \beta_1}{S/\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} \sim t_{n-2}.$$

Proof.

$$\frac{\hat{\beta}_1 - \beta_1}{S/\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} = \frac{\frac{\hat{\beta}_1 - \beta_1}{\sigma/\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}}{\sqrt{\frac{\frac{(n-2)S^2}{\sigma^2}}{n-2}}} \sim t_{n-2},$$

since $\frac{(n-2)S^2}{\sigma^2} \sim \chi^2_{n-2}$ and $\frac{\hat{\beta}_1 - \beta_1}{\sigma/\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} \sim N(0,1)$ are independent. □

Note 26.2. This last theorem provides the key fact needed to test hypotheses or find confidence intervals for β_1.

Example 26.1. Are you more likely to bench press a heavy weight if you can bench press 60lbs many times?

57 female athletes were tested for

$x = \#$(times they could bench press 60lbs) and

$y =$ largest weight they could bench press once.

The data yielded $\frac{S}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} \approx 0.15$ and $\beta_{1,e} = 1.49$

We assume the data follow the linear model and test at the 5% level

$H_0 : \beta_1 = 0$ (meaning x has no influence on Y)

against

$H_1 : \beta_1 > 0$.

A one-sided test is indicated, as it is reasonable to assume that the ability to bench-press 60lbs many times is an indication of strength and wold suggest an ability to lift a heavy weight rather than the inability to do so.

We compute the p-value:

$$P\left(\frac{\hat{\beta}_1 - \beta_1}{S/\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} \geq \frac{1.49 - \beta_1}{0.150} | \beta_1 = 0\right) \approx P(T \geq 9.96) = 0.0000\ldots,$$

where $T \sim t_{55}$, so we reject H_0 and conclude there is strong evidence that the ability to bench press 60lbs many times has a positive effect on the maximal weight one can lift.

26.3 Inference for β_0

Theorem 26.3.

$$\frac{\hat{\beta}_0 - \beta_0}{S\sqrt{\sum_{i=1}^n x_i^2}/\sqrt{n\sum_{i=1}^n (x_i - \bar{x})^2}} \sim t_{n-2}.$$

Note 26.3. This last theorem provides the key fact needed to test hypotheses or find confidence intervals for β_0.

26.4 Testing Equality of Slopes

Theorem 26.4. If $(x_1, Y_1), \ldots (x_n, Y_n)$ and $(x_1^*, Y_1^*), \ldots (x_m^*, Y_m^*)$ satisfy the assumptions of the simple linear model, then

$$\frac{\hat{\beta}_1 - \hat{\beta}_1^* - (\beta_1 - \beta_1^*)}{S\sqrt{\frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} + \frac{1}{\sum_{i=1}^m (x_i^* - \bar{x}^*)^2}}} \sim t_{n+m-4},$$

where

$$S = \sqrt{\frac{\sum_{i=1}^n (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2 + \sum_{i=1}^m (Y_i^* - (\hat{\beta}_0^* + \hat{\beta}_1^* x_i))^2}{n + m - 4}}.$$

Note 26.4. See Example 11.3.4 in the textbook for a nice application of this theorem to the question of genetic diversity.

26.5 Inference for $E[Y|x]$

The true value of $E[Y|x]$ is $\beta_0 + \beta_1 x$, so a natural estimator is $\hat{\beta}_0 + \hat{\beta}_1 x =: \hat{Y}$.

Theorem 26.5.

$$\frac{\hat{Y} - (\beta_0 + \beta_1 x)}{S\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}} \sim t_{n-2}.$$

Proof. $E[\hat{Y}] = \beta_0 + \beta_1 x$ (since $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased).

$$\begin{aligned}
\operatorname{Var}(\hat{Y}) &= \operatorname{Var}(\hat{\beta}_0) + x^2 \operatorname{Var}(\hat{\beta}_1) = \sigma^2 \left(\frac{\sum_{i=1}^n x_i^2 + nx^2}{n \sum (x_i - \bar{x})^2} \right) \\
&= \cdots = \sigma^2 \left(\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right).
\end{aligned}$$

Since $\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}$ and $S^2 = \frac{n}{n-2}\hat{\sigma}^2$, we have $\frac{(n-2)S^2}{\sigma^2} \sim \chi^2_{n-1}$. Therefore,

$$\frac{\frac{\hat{Y} - (\beta_0 + \beta_1 x)}{\sigma\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}}}{\frac{\frac{(n-2)S^2}{\sigma^2}}{n-2}} \sim t_{n-2}.$$

□

Note 26.5. This last theorem provides the key fact needed to test hypotheses or find confidence intervals for $E[Y|x]$. Note that the larger $x - \bar{x}$, the larger the confidence interval.

26.6 Inference for Future Observations

Theorem 26.6. Suppose (x, Y) is a possible future observation of the linear model for which we have the data $(x_1, Y_1), \ldots (x_n, Y_n)$. Then

$$\frac{\hat{Y} - Y}{S\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}} \sim t_{n-2}.$$

Draw a picture containing the 95% confidence bands for $E[Y|x]$ and for $Y|x$.

Math 4501 (Spring 2018) May 14, 2018
Prof. Christian Beneš

Lecture #27: Sample Correlation

27.1 Covariance and Correlation

We have so far treated x as a non-random variable. This makes sense if x represents, for instance, time. However, in many cases, X could be considered as a random variable, just as Y is. We will, for today, consider both X and Y to be random.

Recall

$$\mathrm{Cov}(X,Y) = E[(X-\mu_X)(Y-\mu_Y)] = E[XY] - E[X]E[Y]$$

is the covariance of X and Y. In particular,

$$\mathrm{Cov}(X,X) = \mathrm{Var}(X)$$

and if X and Y are independent, then $\mathrm{Cov}(X,Y) = 0$ (the converse is not true).

$$\mathrm{Corr}(X,Y) = \rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}$$

is the correlation of X and Y.

Fact 27.1. 1. $|\rho(X,Y)| \leq 1$.

2. $|\rho(X,Y)| = 1 \iff Y = aX + b$ for some $a, b \in \mathbb{R}$.

Definition 27.1. The *sample correlation coefficient* of $(X_1, Y_1), \ldots, (X_n, Y_n)$ is

$$R = \frac{\frac{1}{n}\sum_{i=1}^n X_i Y_i - \bar{X}\bar{Y}}{\sqrt{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2}\sqrt{\frac{1}{n}\sum_{i=1}^n (Y_i - \bar{Y})^2}}.$$

It is an estimator for $\rho(X,Y)$.

A useful interpretation of R can be obtained from the following fact:

Proposition 27.1. If $(x_1, y_1), \ldots (x_n, y_n)$ have sample coefficient r, then

$$r = \beta_{1,e}\frac{\sqrt{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}}{\sqrt{n\sum_{i=1}^n y_i^2 - (\sum_{i=1}^n y_i)^2}} = \beta_{1,e}\frac{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}},$$

where $\beta_{1,e}$ is the maximum likelihood estimate for the slope in the linear model.

Proof.

$$\begin{aligned} r &= \frac{n\sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{\sqrt{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}\sqrt{n\sum_{i=1}^n y_i^2 - (\sum_{i=1}^n y_i)^2}} \\ &= \frac{n\sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}\frac{\sqrt{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}}{\sqrt{n\sum_{i=1}^n y_i^2 - (\sum_{i=1}^n y_i)^2}}. \end{aligned}$$

The first fraction in this last product is $\beta_{1,e}$. □

Corollary 27.1. $r > 0 \iff \beta_{1,e} > 0$ and $r < 0 \iff \beta_{1,e} < 0$.

Proposition 27.2.

$$r^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2 - \sum_{i=1}^n (y_i - \beta_{0,e} - \beta_{1,e}x_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}.$$

Corollary 27.2. If $\sum_{i=1}^n (y_i - \beta_{0,e} - \beta_{1,e}x_i)^2 = 0$, that is, if for all $i = 1, \ldots, n$, $y_i = \beta_{0,e} - \beta_{1,e}x_i$, then $r = \pm 1$.

Note 27.1. This last proposition suggests that the more y_i departs from $\beta_{0,e} + \beta_{1,e}x_i$, the smaller $|r|$ will be.

27.2 The Bivariate Normal

Definition 27.2. If $\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho$ are constants, then if X and Y have joint p.d.f.

$$f_{(X,Y)}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2}\frac{1}{1-\rho^2}\left(\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right)\right)$$

then (X, Y) have the *bivariate normal distribution.*

Theorem 27.1. If (X, Y) is bivariate normal as defined above, then

1. The marginals X and Y satisfy $X \sim N(\mu_X, \sigma_X^2), Y \sim N(\mu_Y, \sigma_Y^2)$.
2. $\rho = \text{Corr}(X, Y)$.
3. $E[Y|X = x] = \mu_Y + \rho\frac{\sigma_X}{\sigma_Y}(x - \mu_X)$.
4. $\text{Var}(Y|X = x) = (1 - \rho^2)\sigma_Y^2$.

Note 27.2. For $X \sim N(\mu_X, \sigma_X^2), Y \sim N(\mu_Y, \sigma_Y^2)$, there are infinitely many bivariate normal distributions having X and Y as marginals. However, if ρ is also specified, there is only one.

The following is a very important fact about bivariate normal random variables.

Corollary 27.3. If X and Y are bivariate normal and $\rho(X, Y) = 0$, then X and Y are independent. In particular

$$X, Y \text{ are uncorrelated} \iff X, Y \text{ are independent.}$$

27.3 Inference for the Bivariate Normal

Theorem 27.2. If X, Y are bivariate normal, then the Maximum Likelihood Estimators for $\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho$ are, respectively, $\bar{X}, \bar{Y}, \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2, \frac{1}{n}\sum_{i=1}^n (Y_i - \bar{Y})^2$, and R.

Theorem 27.3. If $(X_1, Y_1), \ldots (X_n, Y_n)$ are a sample from a bivariate normal and R is the sample correlation coefficient, then under $H_0 : \rho = 0$,

$$\frac{\sqrt{n-2}R}{\sqrt{1-R^2}} \sim t_{n-2}.$$

Note 27.3. This theorem is incorrect in some versions of the textbook.

Example 27.1. In the bench-press example from last time, $r \approx 0.8$. We can use this to test $H_0 : r = 0$ against $H_1 : r > 0$ at the 5% significance level. A one-sided test is advisable for the same reason as when we last performed a test for this example. Since $n = 57$ and $r \approx 0.8$, we see that

$$\frac{\sqrt{n-2}r}{\sqrt{1-r^2}} \approx \frac{\sqrt{55}0.8}{\sqrt{1-0.8^2}} \approx 9.88 >> t_{55,0.05},$$

so we again reject H_0 and conclude that there is strong evidence that the number of times one is able to bench-press 60lbs is positively correlated with the largest weight one is able to bench-press once.

Math 4501 (Spring 2018) May 16, 2018
Prof. Christian Beneš

Lecture #28: Multidimensional linear Regression

28.1 Another Look at Linear Regression

Recall that in the simple linear model

$$Y = \beta_0 + \beta_1 x + \epsilon,$$

we found the following maximum likelihood estimators for β_0 and β_1:

$$\hat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^n x_i Y_i - \bar{x}\bar{Y}}{\frac{1}{n}\sum_{i=1}^n x_i^2 - \bar{x}^2}, \qquad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x} = \bar{Y} - \bar{x}\frac{\frac{1}{n}\sum_{i=1}^n x_i Y_i - \bar{x}\bar{Y}}{\frac{1}{n}\sum_{i=1}^n x_i^2 - \bar{x}^2}.$$

Note that we can write this in vector form as follows: If

$$\mathbf{Y} = (Y_1, Y_2, \dots Y_{n-1}, Y_n)' = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_{n-1} \\ Y_n \end{pmatrix}, \quad \mathbf{x} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \\ 1 & x_{n-1} \\ 1 & x_n \end{pmatrix}, \quad \hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)',$$

then we have

$$\mathbf{x}'\mathbf{x}\hat{\beta} = \mathbf{x}'\mathbf{Y},$$

so that

$$\hat{\beta} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{Y},$$

Example 28.1. We will look for the least squres line through the points $(-2, 0), (-1, 0), (0, 1), (1, 1), (2, 3)$. In this problem,

$$\mathbf{y} = (0, 0, 1, 1, 3)', \quad \mathbf{x} = \begin{pmatrix} 1 & -2 \\ 1 & -1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{pmatrix}.$$

Therefore,

$$\mathbf{x}'\mathbf{x} = \begin{pmatrix} 5 & 0 \\ 0 & 10 \end{pmatrix} \Rightarrow (\mathbf{x}'\mathbf{x})^{-1} = \begin{pmatrix} 1/5 & 0 \\ 0 & 1/10 \end{pmatrix}, \quad \mathbf{x}'\mathbf{y} = (5, 7)'.$$

Therefore,

$$\beta_e = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y} = \begin{pmatrix} 1/5 & 0 \\ 0 & 1/10 \end{pmatrix}\begin{pmatrix} 5 \\ 7 \end{pmatrix} = \begin{pmatrix} 1 \\ 7/10 \end{pmatrix}.$$

Therefore,

$$\beta_{0,e} = 1, \quad \beta_{1,e} = 7/10.$$

We can check that R gives the same values (of course it should!):

```
> X=c(-2,-1,0,1,2)
> Y=c(0,0,1,1,3)
> lm.r=lm(Y~X)
> lm.r
Call:
lm(formula = Y ~ X)
Coefficients:
(Intercept) X
1.0 0.7
```

Both ways of doing this (manually or with the help of R) show that the simple linear model for this data set is

$$Y = 1 + 0.7x + \epsilon.$$

Note that we can also use R to do a quadratic regression as follows:

```
> X=c(-2,-1,0,1,2)
> Y=c(0,0,1,1,3)
> lm2.R=lm(Y~X+I(X^2))
> lm2.R
Call:
lm(formula = Y ~ X + I(X^2))
Coefficients:
(Intercept) X I(X^2)
0.5714 0.7000 0.2143
```

This shows that the quadratic model for this data set is

$$Y = 0.5714 + 0.7x + 0.2143x^2 + \epsilon.$$

28.2 Multidimensional Linear Regression

It turns out that the expression above is valid for multidimensional linear regression as well (as long as we re-define $\mathbf{x}, \mathbf{Y}$, and β). Suppose we have the linear model

$$Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \epsilon,$$

where $\epsilon \sim N(0, \sigma^2)$. Define $\mathbf{Y}$ as above and let

$$\mathbf{x} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}, \quad \hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k)'.$$

Then one can show that the maximum likelihood estimator β satisfies

$$\hat{\beta} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{Y}.$$

29 Examples

We will study the relationship between tar content, nicotine content and CO_2 emissions of cigarettes (see http://http://www.amstat.org/publications/jse/datasets/cigarettes.txt for details on the data set). First we download the data:

```
> www=“http://www.amstat.org/publications/jse/datasets/cigarettes.dat”
> CIG=read.table(www)
> CIG
V1 V2 V3 V4 V5
1 Alpine 14.1 0.86 0.9853 13.6
2 Benson&Hedges 16.0 1.06 1.0938 16.6
3 BullDurham 29.8 2.03 1.1650 23.5
4 CamelLights 8.0 0.67 0.9280 10.2
5 Carlton 4.1 0.40 0.9462 5.4
6 Chesterfield 15.0 1.04 0.8885 15.0
7 GoldenLights 8.8 0.76 1.0267 9.0
8 Kent 12.4 0.95 0.9225 12.3
9 Kool 16.6 1.12 0.9372 16.3
10 L&M 14.9 1.02 0.8858 15.4
11 LarkLights 13.7 1.01 0.9643 13.0
12 Marlboro 15.1 0.90 0.9316 14.4
13 Merit 7.8 0.57 0.9705 10.0
14 MultiFilter 11.4 0.78 1.1240 10.2
15 NewportLights 9.0 0.74 0.8517 9.5
16 Now 1.0 0.13 0.7851 1.5
17 OldGold 17.0 1.26 0.9186 18.5
18 PallMallLight 12.8 1.08 1.0395 12.6
19 Raleigh 15.8 0.96 0.9573 17.5
20 SalemUltra 4.5 0.42 0.9106 4.9
21 Tareyton 14.5 1.01 1.0070 15.9
22 True 7.3 0.61 0.9806 8.5
23 ViceroyRichLight 8.6 0.69 0.9693 10.6
```

```
24 VirginiaSlims 15.2 1.02 0.9496 13.9
25 WinstonLights 12.0 0.82 1.1184 14.9
```

We now define the appropriate variables, calling Y the carbon monoxide content (in mg), $X1$ the tar content (in mg), and $X2$ the nicotine content (in mg):

```
> Y=CIG[,5]
> X1=CIG[,2]
> X2=CIG[,3]
```

This now allows us to find a linear model for Y in terms of the variables X1 and X2:

```
> lm.r=lm(Y~ X1+X2)
> lm1.r=lm(Y~X1)
> lm2.r=lm(Y~X2)
> summary(lm.r)
Call:
lm(formula = Y ~ X1 + X2)
Residuals:
Min 1Q Median 3Q Max
-2.899405 -0.784700 -0.001444 0.915854 2.430645
Coefficients:
Estimate Std. Error t value Pr(> |t|)
(Intercept) 3.0896 0.8438 3.662 0.001371 **
X1 0.9625 0.2367 4.067 0.000512 ***
X2 -2.6463 3.7872 -0.699 0.492035
—
Signif. codes: 0 0.001 0.01 0.05 0.1 1
Residual standard error: 1.413 on 22 degrees of freedom
Multiple R-squared: 0.9186, Adjusted R-squared: 0.9112
F-statistic: 124.1 on 2 and 22 DF, p-value: 1.042e-12
> summary(lm1.r)
Call:
lm(formula = Y ~ X1)
Residuals:
Min 1Q Median 3Q Max
-3.1124 -0.7166 -0.3754 1.0091 2.5450
Coefficients:
```

```
Estimate Std. Error t value Pr(> |t|)
(Intercept) 2.74328 0.67521 4.063 0.000481 ***
X1 0.80098 0.05032 15.918 6.55e-14 ***
—
Signif. codes: 0 0.001 0.01 0.05 0.1 1
Residual standard error: 1.397 on 23 degrees of freedom
Multiple R-squared: 0.9168, Adjusted R-squared: 0.9132
F-statistic: 253.4 on 1 and 23 DF, p-value: 6.552e-14
> summary(lm2.r)
Call:
lm(formula = Y ~ X2)
Residuals:
Min 1Q Median 3Q Max
-3.3273 -1.2228 0.2304 1.2700 3.9357
Coefficients:
Estimate Std. Error t value Pr(> |t|)
(Intercept) 1.6647 0.9936 1.675 0.107
X2 12.3954 1.0542 11.759 3.31e-11 ***
—
Signif. codes: 0 0.001 0.01 0.05 0.1 1
Residual standard error: 1.828 on 23 degrees of freedom
Multiple R-squared: 0.8574, Adjusted R-squared: 0.8512
F-statistic: 138.3 on 1 and 23 DF, p-value: 3.312e-11
```