Mathematical Statistics (NYU, Spring 2003)
Summary (answers to his potential exam questions)
By Rebecca Sela

1 Sufficient statistic theorem (1)

Let $X_1, ..., X_n$ be a sample from the distribution $f(x, \theta)$. Let $T(X_1, ..., X_n)$ be a sufficient statistic for θ with continuous factor function $F(T(X_1, ..., X_n), \theta)$. Then,

$$\begin{aligned} P(\vec{X} &\in A|T(\vec{X}) = t) = \lim_{h \to 0} P(\vec{X} \in A| \left|(T(\vec{X}) - t\right| \leq h) \\ &= \lim_{h \to 0} \frac{P(\vec{X} \in A, \left|(T(\vec{X}) - t\right| \leq h)/h}{P(\left|T(\vec{X}) - t\right| \leq h)/h} \\ &= \frac{\frac{d}{dt} P(\vec{X} \in A, T(\vec{X}) \leq t)}{\frac{d}{dt} P(T(\vec{X}) \leq t)} \end{aligned}$$

Consider first the numerator:

$$\begin{aligned} \frac{d}{dt} P(\vec{X} &\in A, T(\vec{X}) \leq t) = \frac{d}{dt} \int_{A \cap \{\vec{x}: T(\vec{x}) = t\}} f(x_1, \theta) ... f(x_n, \theta) dx_1 ... dx_n \\ &= \frac{d}{dt} \int_{A \cap \{\vec{x}: T(\vec{x}) = t\}} F(T(\vec{x}), \theta), h(\vec{x}) dx_1 ... dx_n \\ &= \lim_{h \to 0} \frac{1}{h} \int_{A \cap \{\vec{x}: |T(\vec{x}) - t| \leq h\}} F(T(\vec{x}), \theta), h(\vec{x}) dx_1 ... dx_n \end{aligned}$$

Since $\min_{s \in [t, t+h]} F(s, \theta) \leq F(t, \theta) \leq \max_{s \in [t, t+h]}$ on the interval $[t, t+h]$, we find:

$$\begin{aligned} \lim_{h \to 0} (\min_{s \in [t, t+h]} F(s, \theta)) \frac{1}{h} \int_{A \cap \{\vec{x}: \|T(\vec{x}) - t\| \leq h\}} h(\vec{x}) d\vec{x} &\leq \lim_{h \to 0} \frac{1}{h} \int_{A \cap \{\vec{x}: \|T(\vec{x}) - t\| \leq h\}} F(T(\vec{x}), \theta) h(\vec{x}) d\vec{x} \\ &\leq \lim_{h \to 0} (\max_{s \in [t, t+h]} F(s, \theta)) \frac{1}{h} \int_{A \cap \{\vec{x}: \|T(\vec{x}) - t\| \leq h\}} h(\vec{x}) d\vec{x} \end{aligned}$$

By the continuity of $F(t, \theta)$, $\lim_{h \to 0} (\min_{s \in [t, t+h]} F(s, \theta)) \frac{1}{h} \int_{A \cap \{\vec{x}: \|T(\vec{x}) - t\| \leq h\}} h(\vec{x}) d\vec{x} = \lim_{h \to 0} (\max_{s \in [t, t+h]} F(s, \theta)) \frac{1}{h} \int_{A \cap \{\vec{x}: \|T(\vec{x}) - t\| \leq h\}} h(\vec{x}) d\vec{x} = F(t, \theta)$. Thus,

$$\begin{aligned} \lim_{h \to 0} \frac{1}{h} \int_{A \cap \{\vec{x}: |T(\vec{x}) - t| \leq h\}} F(T(\vec{x}), \theta), h(\vec{x}) dx_1 ... dx_n &= F(t, \theta) \lim_{h \to 0} \frac{1}{h} \int_{A \cap \{\vec{x}: |T(\vec{x}) - t| \leq h\}} h(\vec{x}) d\vec{x} \\ &= F(t, \theta) \frac{d}{dt} \int_{A \cap \{\vec{x}: T(\vec{x}) \leq t\}} h(\vec{x}) d\vec{x} \end{aligned}$$

If we let A be all of R^n, then we have the case of the denominator. Thus, we find:

$$\begin{aligned} P(\vec{X} &\in A|T(\vec{X})=t) = \frac{F(t,\theta)\frac{d}{dt}\int_{A\cap\{\vec{x}:T(\vec{x})\leq t\}} h(\vec{x})d\vec{x}}{F(t,\theta)\frac{d}{dt}\int_{\{\vec{x}:T(\vec{x})\leq t\}} h(\vec{x})d\vec{x}} \\ &= \frac{\frac{d}{dt}\int_{A\cap\{\vec{x}:T(\vec{x})\leq t\}} h(\vec{x})d\vec{x}}{\frac{d}{dt}\int_{\{\vec{x}:T(\vec{x})\leq t\}} h(\vec{x})d\vec{x}} \end{aligned}$$

which is not a function of θ.

Thus, $P(\vec{X} \in A|T(\vec{X})=t)$ does not depend on θ when $T(\vec{X})$ is a sufficient statistic.

2 Examples of sufficient statistics (2)

2.1 Uniform

Suppose $f(x,\theta) = \frac{1}{\theta}I_{(0,\theta)}(x)$. Then,

$$\begin{aligned} \prod f(x_i,\theta) &= \frac{1}{\theta^n}\prod I_{(0,\theta)}(X_i) \\ &= \frac{1}{\theta^n}I_{(-\infty,\theta)}(\max X_i)I_{(0,\infty)}(\min X_i) \end{aligned}$$

Let $F(\max X_i,\theta) = \frac{1}{\theta^n}I_{(-\infty,\theta)}(\max X_i)$ and $h(X_1,...,X_n) = I_{(0,\infty)}(\min X_i)$. This is a factorization of $\prod f(x_i,\theta)$, so $\max X_i$ is a sufficient statistic for the uniform distribution.

2.2 Binomial

Suppose $f(x,\theta) = \theta^{x_i}(1-\theta)^{1-x_i}, x = 0,1$. Then, $\prod f(x_i,\theta) = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}$. Let $T(x_1,...,x_n) = \sum X_i$, $F(t,\theta) = \theta^t(1-\theta)^{n-t}$, and $h(x_1,...,x_n) = 1$. This is a factorization of $\prod f(x_i,\theta)$, which shows that $T(x_1,...,x_n) = \sum X_i$ is a sufficient statistic.

2.3 Normal

Suppose $f(x,\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$. Then,

$$\begin{aligned} \prod f(x_i,\mu,\sigma^2) &= (2\pi\sigma^2)^{-n/2}e^{-\frac{1}{2\sigma^2}\sum(x_i-\mu)^2} \\ &= (2\pi\sigma^2)^{-n/2}e^{-\frac{1}{2\sigma^2}\sum(x_i-\bar{x})^2}e^{-\frac{1}{2\sigma^2}n(\bar{x}-\mu)^2} \end{aligned}$$

since

$$\begin{aligned}
\sum(x_i-\bar{x})^2+n(\bar{x}-\mu) &= \sum(x_i-2x_i\bar{x}+\bar{x}^2)+n(\bar{x}^2-2\mu\bar{x}+\mu^2)\\
&= \sum x_i^2-2\bar{x}(n\bar{x})+n\bar{x}^2+n\bar{x}^2-2n\mu\bar{x}+n\mu^2\\
&= \sum x_i^2-2\mu\sum x_i+n\mu^2\\
&= \sum(x_i^2-2\mu x_i+\mu^2)\\
&= \sum(x_i-\mu)^2
\end{aligned}$$

Case 1: σ^2 unknown, μ known.

Let $T(x_1,...,x_n)=\sum(x_i-\mu)^2$, $F(t,\sigma^2)=(2\pi\sigma^2)^{-n/2}e^{-\frac{1}{2\sigma^2}t}$, and $h(x_1,...,x_n)=1$. This is a factorization of $\prod f(x_i,\sigma^2)$.

Case 2: σ^2 known, μ unknown.

Let $T(x_1,...x_n)=\bar{x}$, $F(t,\mu)=e^{-\frac{1}{2\sigma^2}n(\bar{x}-\mu)^2}$, and $h(x_1,...x_n)=(2\pi\sigma^2)^{-n/2}e^{-\frac{1}{2\sigma^2}\sum(x_i-\bar{x})^2}$. This is a factorization of $\prod f(x_i,\mu)$.

Case 3: μ unknown, σ^2 unknown.

Let $T_1(x_1,...x_n)=\bar{x}$, $T_2(x_1,...x_n)=\sum(x_i-\bar{x})^2$, $F(t_1,t_2,\mu,\sigma^2)=(2\pi\sigma^2)^{-n/2}e^{-\frac{1}{2\sigma^2}t_2}e^{-\frac{1}{2\sigma^2}n(t_1-\mu)^2}$, and $h(x_1,...,x_n)=1$. This is a factorization.

3 Rao-Blackwell Theorem (3)

Let $X_1,...,X_n$ be a sample from the distribution $f(x,\theta)$. Let $Y=Y(X_1,...,X_n)$ be an unbiased estimator of θ. Let $T=T(X_1,...,X_n)$ be a sufficient statistics for θ. Let $\varphi(t)=E(Y|T=t)$.

Lemma 1 *$E(E(g(Y)|T))=E(g(Y))$, for all functions g.*

Proof.

$$\begin{aligned}
E(E(g(Y)|T)) &= \int_{-\infty}^{\infty}E(g(Y)|T)f(t)dt\\
&= \int_{-\infty}^{\infty}(\int_{-\infty}^{\infty}g(y)f(y|t)dy)f(t)dt\\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}g(y)f(y,t)dydt\\
&= \int_{-\infty}^{\infty}g(y)(\int_{-\infty}^{\infty}f(y,t)dt)dy\\
&= \int_{-\infty}^{\infty}g(y)f(y)dy\\
&= E(g(y))
\end{aligned}$$

■

Step 1: $\varphi(t)$ does not depend on θ.

$$\varphi(t) = \int_{R^n} y(x_1, ..., x_n) f(x_1, ..., x_n | T(x_1, ..., x_n) = t) dx_1 ... dx_n.$$

Since $T(x_1, ..., x_n)$ is a sufficient statistic, $f(x_1, ..., x_n | T(x_1, ..., x_n)$ does not depend on θ. Since $y(x_1, ..., x_n)$ is an estimator, it is not a function of θ. Thus, the integral of their product over R^n does not depend on θ.

Step 2: $\varphi(t)$ is unbiased.

$$\begin{aligned} E(\varphi(t)) &= E(E(Y|T)) \\ &= E(Y) \\ &= \theta \end{aligned} \tag{1}$$

by the lemma above.

Step 3: $Var(\varphi(t)) \leq Var(Y)$

$$\begin{aligned} Var(\varphi(T)) &= E(E(Y|T)^2) - E(E(Y|T))^2 \\ &\leq E(E(Y^2|T)) - E(Y)^2 \\ &= E(Y^2) - E(Y)^2 \\ &= Var(Y) \end{aligned}$$

Thus, conditioning an unbiased estimator on the sufficient statistic gives a new unbiased estimator with variance at most that of the old estimator.

4 Some properties of the derivative of the log (4)

Let X have the distribution function $f(x, \theta_0)$. Let $Y = \frac{\partial}{\partial \theta} \log f(X, \theta)|_{\theta=\theta_0}$. Notice that, by the chain rule, $\frac{\partial}{\partial \theta} \log f(x, \theta) = \frac{1}{f(x,\theta)} (\frac{\partial}{\partial \theta} f(x, \theta))$. Using this fact, we find:

$$\begin{aligned}
E(Y) &= E(\frac{\partial}{\partial\theta}\log f(X,\theta)|_{\theta=\theta_0}) \\
&= \int_{-\infty}^{\infty}\frac{\partial}{\partial\theta}\log f(X,\theta)|_{\theta=\theta_0} f(X,\theta_0)dX \\
&= \int_{-\infty}^{\infty}\frac{1}{f(x,\theta_0)}(\frac{\partial}{\partial\theta}f(x,\theta)|_{\theta=\theta_0})f(X,\theta_0)dX \\
&= \int_{-\infty}^{\infty}\frac{\partial}{\partial\theta}f(x,\theta)|_{\theta=\theta_0}dX \\
&= \frac{\partial}{\partial\theta}(\int_{-\infty}^{\infty}f(x,\theta)dX)|_{\theta=\theta_0} \\
&= \frac{\partial}{\partial\theta}(1)|_{\theta=\theta_0} \\
&= 0
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2}{\partial\theta^2}\log f(x,\theta) &= \frac{\partial}{\partial\theta}(\frac{1}{f(x,\theta)}(\frac{\partial}{\partial\theta}f(x,\theta))) \\
&= \frac{1}{f(x,\theta)^2}(f(x,\theta)\frac{\partial^2}{\partial\theta^2}f(x,\theta) - (\frac{\partial}{\partial\theta}f(x,\theta))^2) \\
&= \frac{1}{f(x,\theta)}\frac{\partial^2}{\partial\theta^2}f(x,\theta) - (\frac{1}{f(x,\theta)}\frac{\partial}{\partial\theta}f(x,\theta))^2 \\
&= \frac{1}{f(x,\theta)}\frac{\partial^2}{\partial\theta^2}f(x,\theta) - (\frac{\partial}{\partial\theta}\log f(x,\theta))^2
\end{aligned}$$

$$\begin{aligned}
E(\frac{\partial^2}{\partial\theta^2}\log f(x,\theta)|_{\theta=\theta_0}) &= \int_{-\infty}^{\infty}\frac{1}{f(x,\theta)}\frac{\partial^2}{\partial\theta^2}f(x,\theta)|_{\theta=\theta_0} - (\frac{\partial}{\partial\theta}\log f(x,\theta)|_{\theta=\theta_0})^2dx \\
&= \int_{-\infty}^{\infty}\frac{1}{f(x,\theta)}\frac{\partial^2}{\partial\theta^2}f(x,\theta)|_{\theta=\theta_0}dx - \int_{-\infty}^{\infty}(\frac{\partial}{\partial\theta}\log f(x,\theta)|_{\theta=\theta_0})^2dx \\
&= \frac{\partial^2}{\partial\theta^2}(\int_{-\infty}^{\infty}\frac{1}{f(x,\theta)}f(x,\theta)dx)|_{\theta=\theta_0} - E((\frac{\partial}{\partial\theta}\log f(x,\theta)|_{\theta=\theta_0})^2) \\
&= \frac{\partial^2}{\partial\theta^2}(1)|_{\theta=\theta_0} - E((\frac{\partial}{\partial\theta}\log f(x,\theta)|_{\theta=\theta_0})^2) \\
&= -E((\frac{\partial}{\partial\theta}\log f(x,\theta)|_{\theta=\theta_0})^2)
\end{aligned}$$

Thus, the expected value of Y is zero, and the variance of Y is $-E(\frac{\partial^2}{\partial\theta^2}\log f(x,\theta)|_{\theta=\theta_0})$, which is defined as the information function, $I(\theta)$.

5 The Cramer-Rao lower bound (5)

Let T be an unbiased estimator based on a sample $\vec{X}$, from the distribution $f(x, \theta)$. Then, $E(T) = \theta$. We take the derivative of this equation to find:

$$
\begin{aligned}
1 &= \frac{\partial}{\partial\theta} E(T) && (2)\\
&= \frac{\partial}{\partial\theta} \int_{R^n} T(\vec{x}) f(\vec{x}, \theta) d\vec{x} && (3)\\
&= \int_{R^n} T(\vec{x}) \frac{\partial}{\partial\theta} f(\vec{x}, \theta) d\vec{x} && (4)\\
&= \int_{R^n} T(\vec{x}) (\frac{\partial}{\partial\theta} \log f(\vec{x}, \theta)) f(\vec{x}, \theta) d\vec{x} && (5)\\
&= E(T(\vec{X}) \frac{\partial}{\partial\theta} \log f(\vec{X}, \theta)) && (6)\\
&= E(T(\vec{X}) \frac{\partial}{\partial\theta} \log f(\vec{X}, \theta)) - cE(\frac{\partial}{\partial\theta} \log f(\vec{X}, \theta)) && (7)\\
&= E((T(\vec{X}) - c) \frac{\partial}{\partial\theta} \log f(\vec{X}, \theta)) && (8)\\
&= E((T(\vec{X}) - \theta) \frac{\partial}{\partial\theta} \log f(\vec{X}, \theta)) && (9)
\end{aligned}
$$

By the Cauchy-Schwartz Inequality, $E(AB)^2 \leq E(A^2)E(B^2)$. Squaring both sides of the equation above and applying this, we find:

$$
\begin{aligned}
1 &= E((T(\vec{X}) - \theta) \frac{\partial}{\partial\theta} \log f(\vec{X}, \theta))^2 \\
&\leq E((T(\vec{X}) - \theta)^2) E((\frac{\partial}{\partial\theta} \log f(\vec{X}, \theta))^2) \\
&= Var(T) E((\frac{\partial}{\partial\theta} \log f(\vec{X}, \theta))^2)
\end{aligned}
$$

Since the sample is independent and identically distributed,

$$
\begin{aligned}
(\frac{\partial}{\partial\theta} \log(f(\vec{X}, \theta)))^2 &= (\frac{\partial}{\partial\theta} \log(\prod_{i=1}^{n} f(x_i, \theta)))^2 \\
&= (\frac{\partial}{\partial\theta} \sum_{i=1}^{n} \log f(x_i, \theta))^2 \\
&= (\sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f(x_i, \theta))^2
\end{aligned}
$$

$$
\begin{aligned}
E((\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(x_i,\theta))^2) &= Var(\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(x_i,\theta)) \\
&= \sum_{i=1}^{n} Var(\frac{\partial}{\partial\theta}\log f(x_i,\theta)) \\
&= n\cdot Var(\frac{\partial}{\partial\theta}\log f(x_i,\theta)) \\
&= n\cdot E((\frac{\partial}{\partial\theta}\log f(x_i,\theta))^2) \\
&= nI(\theta)
\end{aligned}
$$

Thus, $1 \leq Var(T)(nI(\theta))$ and the variance of an unbiased estimator is at least $\frac{1}{nI(\theta)}$.

6 Where the Cramer-Rao lower bound fails to hold (6)

Let X be distributed uniform on $(0,\theta)$. That is, $f(x,\theta) = \frac{1}{\theta}I_{(0,\theta)}(x)$. Let $Y = \max X_i$. Then,

$$
\begin{aligned}
P(Y &\leq y) = P(X_1,...X_n \leq y)I_{(0,\theta)}(y) \\
&= (\prod P(X_i \leq y))I_{(0,\theta)}(y) \quad (10) \\
&= (\prod \frac{y}{\theta})I_{(0,\theta)}(y) \quad (11) \\
&= \frac{y^n}{\theta^n}I_{(0,\theta)}(y) \quad (12)
\end{aligned}
$$

$$
\begin{aligned}
f(y) &= \frac{d}{dy}P(Y \leq y) \\
&= n\frac{y^{n-1}}{\theta^n}I_{(0,\theta)}(y)
\end{aligned}
$$

$$
\begin{aligned}
E(Y) &= \int_{-\infty}^{\infty} yf(y)dy \\
&= \int_{0}^{\theta} n\frac{y^n}{\theta^n}dy \\
&= \frac{n}{n+1}\theta
\end{aligned}
$$

$$\begin{aligned}
E(Y^2) &= \int_{-\infty}^{\infty} y^2 f(y) dy \\
&= \int_0^\theta n \frac{y^{n+1}}{\theta^n} dy \\
&= \frac{n}{n+2}\theta^2
\end{aligned}$$

$$\begin{aligned}
Var(Y) &= E(Y^2) - E(Y)^2 \\
&= \frac{n}{n+2}\theta^2 - (\frac{n}{n+1}\theta)^2 \\
&= \theta^2 \frac{n^3 + 2n^2 + n - n^3 - 2n^2}{(n+2)(n+1)^2} \\
&= \theta^2 \frac{n}{(n+2)(n+1)^2}
\end{aligned}$$

Since $E(Y) = \frac{n}{n+1}\theta$, $E(\frac{n+1}{n}Y) = \theta$, and $\frac{n+1}{n} \max X_i$ is an unbiased estimator of θ. The variance of this estimator is given by

$$\begin{aligned}
Var(\frac{n+1}{n}Y) &= \theta^2 \frac{n}{(n+2)(n+1)^2}(\frac{n+1}{n})^2 \\
&= \theta^2 \frac{1}{n(n+2)}
\end{aligned}$$

which is of order $\frac{1}{n^2}$ (and would therefore violate the Cramer-Rao lower bound if it applied).

7 Maximum likelihood estimators for various distributions (7)

7.1 *Normal*

$$\prod_{i=1}^n f(x_i, \mu, \sigma^2) = (\frac{1}{2\pi\sigma^2})^{n/2} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2}$$

$$\log f(x_1, ..., x_n, \mu, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2$$

$$\begin{aligned}
\frac{\partial}{\partial\mu} \log f(x_1, ..., x_n, \mu, \sigma^2) &= -\frac{1}{\sigma^2}\sum_{i=1}^n (x_i - \mu) \\
&= -\frac{n}{\sigma^2}(\bar{x} - \mu)
\end{aligned}$$

$$\begin{aligned}\frac{\partial}{\partial\sigma^2}\log f(x_1,...,x_n,\mu,\sigma^2) &= -\frac{n}{2\sigma^2}+\frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i-\mu)^2\\ &= -\frac{1}{2(\sigma^2)^2}(\sum_{i=1}^{n}(x_i-\mu)^2-n\sigma^2)\end{aligned}$$

If μ is unknown, then we set $\frac{\partial}{\partial\mu}\log f(x_1,...,x_n,\mu,\sigma^2)=0$ to find that $\bar{x}$ is the maximum likelihood estimator of μ.

If σ^2 is unknown and μ is known, then we set $\frac{\partial}{\partial\sigma^2}\log f(x_1,...,x_n,\mu,\sigma^2)=0$ and solve to find that the maximum likelihood estimator of σ^2 is $\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$.

If both μ and σ^2 are unknown, then we estimate μ using its maximum likelihood estimator $\bar{x}$ (which does not depend on σ^2). We use this estimate of μ to maximize with respect to σ^2 and find that the maximum likelihood estimator of σ^2 is $\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$.

7.2 Bernoulli

$$f(x,\theta)=\theta^x(1-\theta)^{1-x}, x=0,1$$

$$f(x_1,...,x_n,\theta)=\theta^{\sum x_i}(1-\theta)^{n-\sum x_i}$$

$$\log f(x_1,...,x_n,\theta)=(\sum x_i)\log\theta+(n-\sum x_i)\log(1-\theta)$$

$$\begin{aligned}\frac{\partial}{\partial\theta}\log f(x_1,...,x_n,\theta) &= \frac{\sum x_i}{\theta}-\frac{n-\sum x_i}{1-\theta}\\ &= \frac{(1-\theta)\sum x_i-\theta(n-\sum x_i)}{\theta(1-\theta)}\\ &= \frac{\theta-\bar{x}}{n\theta(1-\theta)}\end{aligned}$$

Setting the derivative equal to zero, we find that the maximum likelihood estimator of θ is $\bar{X}$.

7.3 *Poisson*

$$f(x,\theta)=\frac{1}{x!}\theta^x e^{-\theta}, x=0,1,2,...$$

$$f(x_1,...,x_n,\theta)=e^{-n\theta}\theta^{\sum x_i}(\prod\frac{1}{x_i!})$$

$$\log f(x_1,...,x_n,\theta)=-n\theta+\sum x_i\log\theta-\sum\log(x_i!)$$

$$\begin{aligned}\frac{\partial}{\partial\theta}\log f(x_1,...,x_n,\theta) &= -n+\frac{1}{\theta}\sum x_i \\ &= \frac{n}{\theta}(\bar{x}-\theta)\end{aligned}$$

Thus, the maximum likelihood estimator of θ is $\bar{x}$.

7.4 Uniform

$$f(x,\theta)=\frac{1}{\theta}I_{(0,\theta)}(x)$$

$$\begin{aligned}f(x_1,...,x_n,\theta) &= \frac{1}{\theta^n}\prod_{i=1}^{n} I_{(0,\theta)}(x_i) \\ &= \frac{1}{\theta^n}I_{(0,\theta)}(\max x_i)\end{aligned}$$

Since $\frac{1}{\theta^n}$ is strictly decreasing, we maximize it by choosing the smallest θ such that $\max X_i \leq \theta$. That is, the maximum likelihood estimator of θ is $\max X_i$.

7.5 Gamma (with β unknown)

$$f(x,\alpha,\beta)=\frac{1}{\beta^\alpha\Gamma(\alpha)}x^{\alpha-1}e^{-x/\beta}$$

$$f(x_1,...,x_n,\alpha,\beta)=(\frac{1}{\beta^\alpha\Gamma(\alpha)})^n(\prod_{i=1}^{n}x_i)^{\alpha-1}e^{-\frac{1}{\beta}\sum x_i}$$

$$\log f(x_1,...,x_n,\theta)=-n\alpha\log\beta-n\log\Gamma(\alpha)+(\alpha-1)\sum\log x_i-\frac{\sum x_i}{\beta}$$

$$\begin{aligned}\frac{\partial}{\partial\theta}\log f(x_1,...,x_n,\theta) &= -\frac{n\alpha}{\beta}+\frac{1}{\beta^2}\sum x_i \\ &= \frac{1}{\beta^2}(\sum x_i-n\alpha\beta)\end{aligned}$$

The derivative is 0 for $\beta=\frac{\bar{x}}{\alpha}$. Thus, $\bar{x}/\alpha$ is the maximum likelihood estimator.

8 The likelihood equation will have a solution (8)

Let $x_1, ..., x_n$ be a sample from the distribution $f(x, \theta_0)$. By the Law of Large Numbers, the sequence $\{\frac{1}{n}\sum_{i=1}^n Y_i\}$ converges to $E(Yi)$ with probability 1. In particular, the sequence $\{\frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial\theta}\log f(x_i, \theta)|_{\theta=\theta_0}\}$ converges to $E(\frac{\partial}{\partial\theta}\log f(x, \theta)|_{\theta=\theta_0}) = 0$, and the sequence $\{\frac{1}{n}\sum_{i=1}^n \frac{\partial^2}{\partial\theta^2}\log f(x_i, \theta)|_{\theta=\theta_0}\}$ converges to $E(\frac{\partial^2}{\partial\theta^2}\log f(x, \theta)|_{\theta=\theta_0}) = -I(\theta)$, with probability 1. Let $g_n(\theta) = \frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial\theta}\log f(x_i, \theta)$. Then, $\lim_{n\longrightarrow\infty} g_n(\theta_0) = 0$ and $\lim_{n\to\infty} g_n'(\theta_0) = -I(\theta_0) \neq 0$.

Let $\varepsilon > 0$ be given. By the Taylor expansion,

$$g_n(\theta) \approx g_n(\theta_0) + g_n'(\theta_0)(\theta - \theta_0), |\theta - \theta_0| < \varepsilon$$

For sufficiently large n, then,

$$\begin{aligned} g_n(\theta) &= -I(\theta_0)(\theta - \theta_0) \\ \frac{1}{\theta - \theta_0} g_n'(\theta_0) &= -I(\theta_0) \end{aligned}$$

Since $-I(\theta_0) < 0$ in the interval $(\theta_0 - \varepsilon, \theta_0 + \varepsilon)$ while $\theta - \theta_0$ is both positive and negative in that interval, $g_n(\theta)$ must change sign in this interval as well. Since $g_n(\theta)$ is continuous, this means $\frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial\theta}\log f(x_i, \theta) = g_n(\theta) = 0$ in this interval, and the likelihood equation has a solution in $(\theta - \varepsilon, \theta + \varepsilon)$ with probability one for large n.

9 The limiting distribution of the maximum likelihood estimators (9)

Let $x_1, ..., x_n$ be a sample from the distribution $f(x, \theta_0)$. Recall that $E(\frac{\partial}{\partial\theta}\log f(x, \theta)|_{\theta=\theta_0} = 0)$ and that $Var(\frac{\partial}{\partial\theta}\log f(x, \theta)|_{\theta=\theta_0}) = E((\frac{\partial}{\partial\theta}\log f(x_i, \theta)|_{\theta=\theta_0})^2) = I(\theta_0)$. Thus, by the Central Limit Theorem, $\frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{\partial}{\partial\theta}\log f(x_i, \theta)|_{\theta=\theta_0}$ has a normal limiting distribution with mean 0 and variance $I(\theta_0)$. By the Law of Large Numbers, since $E(\frac{\partial^2}{\partial\theta^2}\log f(x, \theta)|_{\theta=\theta_0}) = -I(\theta_0)$, $\lim_{n\to\infty} \frac{1}{n}\sum_{i-1}^n \frac{\partial^2}{\partial\theta^2}\log f(x, \theta)|_{\theta=\theta_0} = -I(\theta_0)$ with probability one.

Consider $\frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{\partial}{\partial\theta}\log f(x_i, \theta)$. Using the Taylor expansion about θ_0, we find:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{\partial}{\partial\theta}\log f(x_i, \theta) \approx \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{\partial}{\partial\theta}\log f(x_i, \theta)|_{\theta=\theta_0} + \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{\partial^2}{\partial\theta^2}\log f(x_i, \theta)|_{\theta=\theta_0}(\theta - \theta_0)$$

in a neighborhood of θ_0. For large n, there is a solution to the likelihood equation in this neighborhood with probability one. Let $\hat{\theta}_n$ be this solution. Then, we have:

$$\begin{aligned} 0 &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(x_i,\theta)|_{\theta=\hat{\theta}_n} \\ &\approx \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(x_i,\theta)|_{\theta=\theta_0} + \sqrt{n}(\hat{\theta}_n-\theta_0)(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta^2}\log f(x_i,\theta)|_{\theta=\theta_0}) \end{aligned}$$

Substituting in the limit, for $\lim_{n\to\infty}\frac{1}{n}\sum_{i-1}^{n}\frac{\partial^2}{\partial\theta^2}\log f(x,\theta)|_{\theta=\theta_0} = -I(\theta_0)$ and solving for $\sqrt{n}(\hat{\theta}_n-\theta_0)$, we find:

$$\sqrt{n}(\hat{\theta}_n-\theta_0) \approx \frac{1}{I(\theta_0)}(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(x_i,\theta)|_{\theta=\theta_0})$$

The right hand side is the product of $\frac{1}{I(\theta_0)}$ and a normal variable with mean 0 and variance $I(\theta_0)$; that is, a normal variable with mean 0 and variance $\frac{1}{I(\theta_0)^2}I(\theta_0) = \frac{1}{I(\theta_0)^2}$. Thus, $\sqrt{n}(\hat{\theta}_n-\theta_0)\tilde{\ }N(0,\frac{1}{I(\theta_0)})$.

10 Confidence Intervals for Various Distributions (10)

10.1 Normal, σ^2 known:

Let $X\tilde{\ }Normal(\mu,\sigma^2)$, with σ^2 known. Let $\alpha < 1$ be given (this is the confidence level). Set $Z = \frac{X-\mu}{\sigma}$. Then, $Z\tilde{\ }Normal(0,1)$. Choose $z_{\alpha/2}$ such that $\Phi(z_{\alpha/2}) = \int_{-\infty}^{z_{\alpha/2}}\phi(z)dz = 1-\frac{\alpha}{2}$. Then, by symmetry:

$$\begin{aligned} 1-\alpha &= P(-z_{\alpha/2} \le Z \le z_{\alpha/2}) \\ &= P(-z_{\alpha/2} \le \frac{X-\mu}{\sigma} \le z_{\alpha/2}) \\ &= P(-\sigma z_{\alpha/2} \le X-\mu \le \sigma z_{\alpha/2}) \\ &= P(X - z_{\alpha/2}\sigma \le \mu \le X + z_{\alpha/2}\sigma) \end{aligned}$$

and $(X - z_{\alpha/2}\sigma, X + z_{\alpha/2}\sigma)$ is a $1-\alpha$ confidence interval for μ.

If we choose a sample of size n, then $\bar{X}\tilde{\ }Normal(\mu,\frac{\sigma^2}{n})$, and then

$$\begin{aligned} 1-\alpha &= P(-z_{\alpha/2} \le \frac{X-\mu}{\sigma/\sqrt{n}} \le z_{\alpha/2}) \\ &= P(X - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le X + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) \end{aligned}$$

so that $(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}})$ is a $1-\alpha$ confidence interval for μ.

10.2 Normal, σ^2 unknown:

Let $X_1, ..., X_n$ be a sample from the distribution $Normal(\mu, \sigma^2)$. Then, $\bar{X}\tilde{\ } Normal(\mu, \frac{\sigma^2}{n})$ and $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is distributed χ^2 with $n-1$ degrees of freedom, so that $T = \frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{((n-1)\frac{s^2}{\sigma^2})/(n-1)}} = \frac{\bar{X}-\mu}{s/\sqrt{n}}$ has a t-distribution with $n-1$ degrees of freedom. Choose $t_{\alpha/2,n-1}$ such that $P(T \leq t_{\alpha/2,n-1}) = 1 - \frac{\alpha}{2}$. Then, by symmetry,

$$\begin{aligned} 1-\alpha &= P(-t_{\alpha/2,n-1} \leq \frac{\bar{X}-\mu}{s/\sqrt{n}} \leq t_{\alpha/2,n-1}) \\ &= P(\bar{X} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}) \end{aligned}$$

and $(\bar{X} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}, \bar{X} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}})$ is a $1-\alpha$ confidence interval for μ.

If n is sufficiently large, then the t-distribution with $n-1$ degrees of freedom is approximately the standard normal distribution, and $t_{\alpha/2,n-1} \approx z_{\alpha/2.}$

10.3 Binomial:

If X is a Bernoulli random variable (with probability θ of success) then, by the Central Limit Theorem, $\bar{X}$ is distributed approximately normal with mean θ and variance $\theta(1-\theta)$. Since θ is unknown, we must approximate the variance. Note that $\theta(1-\theta) \leq 0.5(1-0.5) = 0.25$ for all values of θ. Thus, 0.25 is a conservative estimate of the variance and we may use normal confidence intervals with this estimate of the variance.

10.4 Non-normal, large sample:

We may use the Central Limit Theorem and the fact that the maximum likelihood estimator is approximately normally distributed with mean θ and variance $I(\theta)$ to construct an approximate confidence interval using the methods above. (We should estimate $I(\theta)$ by $I(\hat{\theta})$ or by some upper bound, as in the binomial case.)

11 The sum of normals squared is chi-squared (11)

Let $X_1, ..., X_n$ be a sample from a standard normal distribution. Consider the cumulative distribution function of $\sum_{i=1}^{n} X_i^2$:

$$\begin{aligned} P(\sum_{i=1}^{n} X_i^2 &\leq y) = \int_{\{x_1,...x_n: \sum x_i^2 \leq y\}} f(x_1, ..., x_n, \theta) dx_1...dx_n \\ &= \int_{\{x_1,...x_n: \sum x_i^2 \leq y\}} (2\pi)^{-n/2} e^{-\sum x_i^2} dx_1...dx_n \end{aligned}$$

Taking the derivative with respect to y, we find:

$$\begin{aligned} f(y) &= \frac{d}{dy} P(\sum_{i=1}^{n} X_i^2 \leq y) \\ &= \frac{d}{dy} \int_{\{x_1,...x_n: \sum x_i^2 \leq y\}} (2\pi)^{-n/2} e^{-\sum x_i^2} dx_1...dx_n \end{aligned}$$

Recall that $\frac{d}{dt} \int_{\{x_1,...,x_n | T(x_1,...,x_n)=t\}} F(T(x_1, ..., x_n), \theta) h(x_1, ..., x_n) dx_1...dx_n = F(t, \theta) \frac{d}{dt} \int_{\{x_1,...,x_n | T(x_1,...,x_n)=t\}} h(x_1, ..., x_n) dx_1...dx_n$. Thus, we find:

$$f(y) = (2\pi)^{-n/2} e^{-y} \frac{d}{dy} \int_{\{x_1,...x_n: \sum x_i^2 \leq y\}} dx_1...dx_n$$

Notice that $\int_{\{x_1,...x_n: \sum x_i^2 \leq y\}} dx_1...dx_n$ is the volume of an n-ball of radius y, which is proportional to $y^{n/2}$. Thus, $\frac{d}{dy} \int_{\{x_1,...x_n: \sum x_i^2 \leq y\}} dx_1...dx_n$ is proportional to $\frac{n}{2} y^{\frac{n}{2}-1}$. Thus, $f(y)$ is proportional to $e^{-y} y^{\frac{n}{2}-1}$, which is the $Gamma(\frac{n}{2}, 1) = \chi_n^2$ distribution. Thus, $\sum_{i=1}^{n} X_i^2$ has a chi-squared distribution with n degrees of freedom.

12 Independence of the estimated mean and standard deviation (12)

Let $X_1, ...X_n$ be a sample from a $Normal(\mu, \sigma^2)$ distribution. Set $Z_i = \frac{X_i - \mu}{\sigma}$. Then, $Z_1, ..., Z_n$ are distributed $Normal(0, 1)$. Let $U = \sqrt{n}\bar{Z}$ and $V = (n-1)s^2 = \sum (Z_i - \bar{Z})^2$. The joint distribution function of these two variables is given by:

$$F(u, v) = \int_{\{Z_i,...Z_n: \sqrt{n}\bar{Z} \leq u, (n-1)s^2 \leq v\}} (2\pi)^{-n/2} e^{-\sum Z_i^2} dZ_1...dZ_n$$

Let P be any real orthonormal matrix with first row $(\frac{1}{\sqrt{n}}, ..., \frac{1}{\sqrt{n}})$; such a matrix exists because this vector is of length one and we may apply the Gram-Schmidt orthogonalization. Set $\vec{Y} = P\vec{Z}$. Then, $Y_1 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i = U$. Since orthogonal matrices preserve inner products, $\sum_{i=1}^{n} Y_i^2 = \sum_{i=1}^{n} Z_i^2$, and $V = \sum_{i=1}^{n} (Z_i - \bar{Z})^2 = \sum_{i=1}^{n} Z_i^2 - n\bar{Z}^2 = \sum_{i=1}^{n} Y_i^2 - Y_1^2 = \sum_{i=2}^{n} Y_i^2$. Substituting $Y_1, ..., Y_n$ for $Z_1, ..., Z_n$ in the joint distribution function above, we find:

$$
\begin{aligned}
F(u,v) &= \int_{\{Y_1,...,Y_n:Y_1\leq u,\sum_{i=2}^n Y_i^2\leq v\}} (2\pi)^{-n/2} e^{\sum_{i=1}^n Y_i^2} dY_1...dY_n \\
&= (2\pi)^{-n/2}(\int_{-\infty}^{u} e^{Y_1^2} dY_1)(\int_{\{Y_2,...,Y_n:\sum_{i=2}^n Y_i^2\leq v\}} e^{\sum_{i=2}^n Y_i^2} dY_2...dY_n)
\end{aligned}
$$

Thus, we see that the density function factors, meaning that U and V are independent. Furthermore, the factor containing $U = \sqrt{n}Z$ is of the form of a standard normal random variable, and the factor containing $V = (n-1)s^2$ is of the form of a χ^2 random variable with $n-1$ degrees of freedom. Since $U = \sqrt{n}\bar{Z} = \sqrt{n}\frac{\bar{x}-\mu}{\sigma}$ and $V = (n-1)s^2 = \sum(\frac{X_i-\mu}{\sigma} - \frac{\bar{X}-\mu}{\sigma})^2 = \frac{\sum(X_i-\bar{X})^2}{\sigma^2} = \frac{s_x^2(n-1)}{\sigma^2}$, we have shown that $\sqrt{n}\frac{\bar{x}-\mu}{\sigma}$ and $\frac{s_x^2(n-1)}{\sigma^2}$ are independent, with the former normally distributed and the latter distributed chi-square.

13 The t-distribution (13)

Let X be a standard normal random variable and Y an independently distributed chi-square variable with n degrees of freedom. Let $T = \frac{X}{\sqrt{y/n}}$. The T is distributed with a t-distribution with n degrees of freedom.

If $X_1, ..., X_n$ are independently distributed normal random variables, then $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ and $\frac{(n-1)s^2}{\sigma^2}$ are independently distributed standard normal and chi-square with $n-1$ degrees of freedom random variables respectively. Thus,

$$
\begin{aligned}
T &= \frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{1}{n-1}(\frac{(n-1)s^2}{\sigma^2})}} \\
&= \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \cdot \frac{\sigma}{s}\sqrt{\frac{n-1}{n-1}} \\
&= \frac{\bar{X}-\mu}{s/\sqrt{n}}
\end{aligned}
$$

has a t-distribution with $n-1$ degrees of freedom.

14 Some facts about hypothesis tests, levels of significance, and power (14)

Let the paramter space be the disjoint union of H_0 and H_1. A hypothesis test is a mapping from a sample $X_1, ..., X_n$ to the set $\{H_0, H_1\}$. The inverse image of H_1 is called the critical region, C. $P_\theta(C) = P(C|\theta)$ is the probability of rejecting H_0 if θ is the true parameter value; this is called the power function. The level of significance, which is the maximum probability of a false rejection, is $\sup_{\theta\in H_0} P_\theta(C)$.

14.1 A simple normal hypothesis test

Suppose X is distributed $Normal(\mu, \sigma^2)$, with σ^2 known. Let $H_0 = \{\mu : \mu \leq \mu_0\}$ and $H_1 = \{\mu : \mu > \mu_0\}$. Let $C_1 = \{x : x > \mu\}$. Then,

$$\begin{aligned} P_\mu(C_1) &= P_\mu(X > \mu_0) \\ &= P_\mu(\frac{X-\mu}{\sigma} > \frac{\mu_0 - \mu}{\sigma}) \\ &= 1 - \Phi(\frac{\mu_0 - \mu}{\sigma}) \end{aligned}$$

Since Φ is an increasing function, the function above is maximized by choosing the largest possible μ is H_0. Therefore, the level of significance is:

$$\begin{aligned} \sup_{\mu \leq \mu_0} P_\mu(C_1) &= 1 - \Phi(\frac{\mu_0 - \mu_0}{\sigma}) \\ &= 1 - \frac{1}{2} = \frac{1}{2} \end{aligned}$$

Consider a second critical region: $C_2 = \{\bar{X} : \bar{X} > \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}\}$. Then,

$$\begin{aligned} P_\mu(C_2) &= P_\mu(\bar{X} > \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}) \\ &= P_\mu(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \geq \frac{\mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}} - \mu}{\sigma/\sqrt{n}}) \\ &= 1 - \Phi(\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} + z_\alpha) \end{aligned}$$

Then the level of significance is:

$$\begin{aligned} \sup_{\mu \leq \mu_0} P_\mu(C_2) &= \sup_{\mu \leq \mu_0} 1 - \Phi(\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} + z_\alpha) \\ &= 1 - \Phi(z_\alpha) \\ &= 1 - \alpha \end{aligned}$$

15 The Neyman-Pearson Lemma (15)

Theorem 2 *Let $\vec{X}$ be a sample from the distribution $f(x;\theta)$. Let H_0 and H_1 be the simple hypotheses $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_1$. Define $R(\vec{X}) = \frac{\prod_{i=1}^n f(x_i, \theta_1)}{\prod_{i=1}^n f(x_i, \theta_0)}$. Define a critical region $C \subset R^n$ for a given $\lambda \in R$ by $C = \{\vec{x} \in R^n : R(\vec{x}) > \lambda\}$; that is, reject when $R(\vec{x}) > \lambda$. If D is the critical region for any test such that*

$P_{\theta_0}(D) \leq P_{\theta_0}(C)$ then $P_{\theta_1}(D) \leq P_{\theta_1}(C)$. (The test based on the likelihood ration is the most powerful for a given level of significance.)

Proof. *For any $A \subset R^n$, define $P_0(A) = \int_A \prod_{i=1}^{n} f(x_i, \theta_0) dx_1 ... dx_n$ and $P_1(A) = \int_A \prod_{i=1}^{n} f(x_i, \theta_1) dx_1 ... dx_n$. Thus, we want to show that $P_0(D) \leq P_0(C)$ implies that $P_1(D) \leq P_1(C)$. Let D be given. Then:*

$$\begin{aligned} P_1(D \cap C^C) &= \int_{D \cap \{\vec{x}: R(\vec{x}) \leq \lambda\}} \prod_{i=1}^{n} f(x_i, \theta_1) dx_1 ... dx_n \\ &= \int_{D \cap \{\vec{x}: R(\vec{x}) \leq \lambda\}} R(\vec{x}) \prod_{i=1}^{n} f(x_i, \theta_0) dx_1 ... dx_n \\ &\leq \int_{D \cap \{\vec{x}: R(\vec{x}) \leq \lambda\}} \lambda \prod_{i=1}^{n} f(x_i, \theta_0) dx_1 ... dx_n \\ &= \lambda P_0(D \cap C^C) \end{aligned}$$

$$\begin{aligned} P_1(D^C \cap C) &= \int_{D^C \cap \{\vec{x}: R(\vec{x}) \geq \lambda\}} \prod_{i=1}^{n} f(x_i, \theta_1) dx_1 ... dx_n \\ &= \int_{D^C \cap \{\vec{x}: R(\vec{x}) \geq \lambda\}} R(\vec{x}) \prod_{i=1}^{n} f(x_i, \theta_0) dx_1 ... dx_n \\ &\geq \int_{D^C \cap \{\vec{x}: R(\vec{x}) \geq \lambda\}} \lambda \prod_{i=1}^{n} f(x_i, \theta_0) dx_1 ... dx_n \\ &= \lambda P_0(D^C \cap C) \end{aligned}$$

Combining these facts, we find:

$$\begin{aligned} P_1(D) &= P_1(D \cap C) + P_1(D \cap C^C) \\ &\leq P_1(D \cap C) + \lambda P_0(D \cap C^C) \\ &= P_1(D \cap C) + \lambda (P_0(D) - P_0(D \cap C)) \\ &\leq P_1(D \cap C) + \lambda (P_0(C) - P_0(D \cap C)) \\ &= P_1(D \cap C) + \lambda (P_0(C \cap D^C)) \\ &\leq P_1(D \cap C) + P_1(C \cap D^C) \\ &= P_1(C) \end{aligned}$$

■

16 The likelihood ratio test for the normal distribution (16)

Suppose $\vec{X}$ is a sample from the distribution $Normal(\mu, \sigma^2)$. Let the null hypothesis be $H_0 : \mu = \mu_0$, σ^2 unknown. Let the alternative hypothesis be

$H_1 : \mu \neq \mu_0$, σ^2 unknown. The likelihood function of this distribution is:

$$\begin{aligned} L(\mu, \sigma^2, \vec{X}) &= \prod_{i=1}^{n} (\frac{1}{2\pi\sigma^2}) e^{-\frac{1}{2\sigma^2}(X_i - \mu)^2} \\ &= (2\pi\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2}\sum(X_i-\mu)^2} \end{aligned}$$

Under the null hypothesis, $\hat{\hat{\mu}} = \mu_0$, so we maximize the likelihood function with respect to σ^2:

$$\log L(\mu, \sigma^2, \vec{X}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum(X_i - \mu_0)^2$$

$$\begin{aligned} \frac{\partial}{\partial\sigma^2}\log L(\mu, \sigma^2, \vec{X}) &= -\frac{n}{\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum(X_i - \mu_0)^2 \\ &= 0 \end{aligned}$$

$$\hat{\hat{\sigma}}^2 = \frac{1}{n}\sum(X_i - \mu_0)^2$$

Under the alternative hypothesis, we have the maximum likelihood estimates: $\hat{\mu} = \bar{X}$ and $\hat{\sigma}^2 = \frac{1}{n}\sum(X_i - \bar{X})^2$.

Then, the likelihood ratio of the null and alternative hypotheses is:

$$\begin{aligned} \lambda &= \frac{L(\hat{\hat{\mu}}, \hat{\hat{\sigma}}^2)}{L(\hat{\mu}, \hat{\sigma}^2)} \\ &= \frac{(2\pi\hat{\hat{\sigma}}^2)^{-n/2} e^{-\frac{1}{2\hat{\hat{\sigma}}^2}\sum(X_i-\hat{\hat{\mu}})^2}}{(2\pi\hat{\sigma}^2)^{-n/2} e^{-\frac{1}{2\hat{\sigma}^2}\sum(X_i-\hat{\mu})^2}} \\ &= (\frac{\hat{\sigma}^2}{\hat{\hat{\sigma}}^2}) e^{-\frac{1}{2\hat{\hat{\sigma}}^2}\sum(X_i-\hat{\hat{\mu}})^2 + \frac{1}{2\hat{\sigma}^2}\sum(X_i-\hat{\mu})^2} \\ &= (\frac{\sum(X_i - \bar{X})^2}{\sum(X_i - \mu_0)^2})^{n/2} e^{-\frac{n}{2}+\frac{n}{2}} \\ &= (\frac{\sum(X_i - \bar{X})^2}{\sum(X_i - \mu_0)^2})^{n/2} \end{aligned}$$

Recall that $\sum(X_i - \mu_0)^2 = \sum(X_i - \bar{X})^2 + n(\bar{X} - \mu_0)^2$, so

$$\begin{aligned} \lambda &= (\frac{\sum(X_i - \bar{X})^2}{\sum(X_i - \bar{X})^2 + n(\bar{X} - \mu_0)^2})^{\frac{n}{2}} \\ &= (\frac{1}{1 + n\frac{(\bar{X}-\mu_0)^2}{\sum(X_i-\bar{X})^2}})^{\frac{n}{2}} \end{aligned}$$

is a decreasing function of $n\frac{(\bar{X}-\mu_0)^2}{\sum(X_i-\bar{X})^2} = (\frac{|\bar{X}-\mu_0|}{\sqrt{n-1}s/\sqrt{n}})^2 = (\frac{t}{\sqrt{n-1}})^2$, which is an increasing function of the t-statistic. Thus, λ is minimized if and only if the t-statistic is large, and a t-test is a likelihood ratio test.

17 Linear combinations of multivariate normals (and MGF's) (17)

17.1 Case 1: Standard Normal

The moment generating function for a standard normal variable, $\vec{Z}$, is given by:

$$\begin{aligned}
E(e^{\vec{t}'\vec{Z}}) &= E(e^{\sum t_j z_j}) \\
&= \prod_{j=1}^{m} E(e^{t_j z_j}) \\
&= \prod_{j=1}^{m} e^{\frac{1}{2}t_j^2} \\
&= e^{\frac{1}{2}\sum t_j^2} \\
&= e^{\frac{1}{2}\|\vec{t}\|^2}
\end{aligned}$$

17.2 Case 2: General normal

Let $\vec{X}$ be a multivariate normal random vector with distribution $Normal_m(\vec{\mu}, \Sigma)$. Then, we may write $\vec{X} = \Sigma^{1/2}\vec{Z} + \vec{\mu}$, where $\vec{\mu}$ is distributed $Normal_m(\vec{0}, I)$. Then, the moment generating function is:

$$\begin{aligned}
E(e^{\vec{t}'\vec{X}}) &= E(e^{\vec{t}'(\Sigma^{1/2}\vec{Z}+\vec{\mu})}) \\
&= E(e^{\vec{t}'\Sigma^{1/2}\vec{Z}}e^{\vec{t}'\vec{\mu}}) \\
&= e^{\vec{t}'\vec{\mu}}E(e^{\vec{t}'\Sigma^{1/2}\vec{Z}}) \\
&= e^{\vec{t}'\vec{\mu}}E(e^{(\Sigma^{1/2}\vec{t})'\vec{Z}}) \\
&= e^{\vec{t}'\vec{\mu}}e^{\frac{1}{2}\|'\Sigma^{1/2}\vec{t}\|^2} \\
&= e^{\vec{t}'\vec{\mu}}e^{\frac{1}{2}(\vec{t}'\Sigma\vec{t})} \\
&= e^{\vec{t}'\vec{\mu}+\frac{1}{2}\vec{t}'\Sigma\vec{t}}
\end{aligned}$$

17.3 Case 3: Linear combinations of normal variables

Let $\vec{Y} = B\vec{X}$, where B is not necessarily symmetric, invertible, or square. Then, $\vec{Y} = B\vec{X} = B\Sigma^{1/2}\vec{Z} + B\vec{\mu}$, and its moment generating function is:

$$\begin{aligned}
E(e^{\vec{t}'\vec{Y}}) &= E(e^{\vec{t}'(B\Sigma^{1/2}\vec{Z}+B\vec{\mu})}) \\
&= e^{\vec{t}'B\vec{\mu}}E(e^{\vec{t}'B\Sigma^{1/2}\vec{Z}}) \\
&= e^{\vec{t}'B\vec{\mu}}e^{\frac{1}{2}\|'\Sigma^{1/2}B'\vec{t}\|^2} \\
&= e^{\vec{t}'B\vec{\mu}}e^{\frac{1}{2}('\Sigma^{1/2}B'\vec{t})'('\Sigma^{1/2}B'\vec{t})} \\
&= e^{\vec{t}'B\vec{\mu}+\frac{1}{2}\vec{t}'B\Sigma B'\vec{t}}
\end{aligned}$$

Thus, $\vec{Y} = B\vec{X}$ is also multivariate normal, with mean $B\mu$ and covariance matrix $B\Sigma B'$.

18 The estimated regression coefficients minimize the estimated errors (18)

Let $\vec{X}$ be a vector of length n and A be an $n \times k$ matrix of rank k. Let $\hat{\vec{\theta}} = (A'A)^{-1}A'\vec{X}$. Then,

$$\begin{aligned}\left\|\vec{X} - A\vec{\theta}\right\|^2 &= \left\|\vec{X} - A\vec{\theta} + A\hat{\vec{\theta}} - A\hat{\vec{\theta}}\right\|^2 \\ &= \left\|\vec{X} - A\hat{\vec{\theta}}\right\|^2 + \left\|A\hat{\vec{\theta}} - A\vec{\theta}\right\|^2 + 2(\vec{X} - A\hat{\vec{\theta}})'(A\hat{\vec{\theta}} - A\vec{\theta})\end{aligned}$$

The last term is zero:

$$\begin{aligned}(\vec{X} - A\hat{\vec{\theta}})'(A\hat{\vec{\theta}} - A\vec{\theta}) &= \vec{X}'A\hat{\vec{\theta}} - \vec{X}'A\vec{\theta} - \hat{\vec{\theta}}'A'A\hat{\vec{\theta}} + \hat{\vec{\theta}}A'A\vec{\theta} \\ &= \vec{X}'A(A'A)^{-1}A'\vec{X} - \vec{X}'A\vec{\theta} - \vec{X}'A(A'A)^{-1}A'A(A'A)^{-1}A'\vec{X} + \vec{X}'A(A'A)^{-1}A'A\vec{\theta} \\ &= \vec{X}'A(A'A)^{-1}A'\vec{X} - \vec{X}'A\vec{\theta} - \vec{X}'A(A'A)^{-1}A'\vec{X} + \vec{X}'A\vec{\theta} \\ &= 0\end{aligned}$$

Thus,

$$\left\|\vec{X} - A\vec{\theta}\right\|^2 = \left\|\vec{X} - A\hat{\vec{\theta}}\right\|^2 + \left\|A\hat{\vec{\theta}} - A\vec{\theta}\right\|^2$$

Since $\vec{X}$, A, and therefore $\hat{\vec{\theta}}$ are all fixed, we minimize the expression above by choosing $\vec{\theta}$ in the second term. Since $\left\|A\hat{\vec{\theta}} - A\vec{\theta}\right\|^2 \geq 0$ for all values of $\vec{\theta}$, we choose $\vec{\theta} = \hat{\vec{\theta}}$, so that $\left\|A\hat{\vec{\theta}} - A\vec{\theta}\right\|^2 = 0$ and $\left\|\vec{X} - A\vec{\theta}\right\|^2$ is minimized.

19 The properties of the least squares coefficients (19)

Suppose $\vec{X}$ is a random vector with mean $A\vec{\theta}$ and covariance matrix $\sigma^2 I$. Let $\hat{\vec{\theta}} = (A'A)^{-1}A'\vec{X}$. Then,

$$\begin{aligned}E(\hat{\vec{\theta}}) &= E((A'A)^{-1}A'\vec{X}) \\ &= (A'A)^{-1}A'E(\vec{X}) \\ &= (A'A)^{-1}A'A\vec{\theta} \\ &= \vec{\theta}\end{aligned}$$

$$\begin{aligned} Cov(\hat{\vec{\theta}}) &= Cov((A'A)^{-1}A'\vec{X}) \\ &= (A'A)^{-1}A'(\sigma^2 I)A(A'A)^{-1} \\ &= \sigma^2(A'A)^{-1} \end{aligned}$$

If $\vec{X}$ is normally distributed, then $\hat{\vec{\theta}}$ is a linear transformation of $\vec{X}$ and thus is normally distributed as well.

20 Estimating the variance of the least squares coefficients (20)

Suppose $\vec{X}$ is a multivariate normal random vector with mean $A\vec{\theta}$ and covariance matrix $\sigma^2 I$. Let $\hat{\vec{\theta}} = (A'A)^{-1}A'\vec{X}$.

Let $V \in R^k$ be any vector. Then,

$$\begin{aligned} & \left\| (\vec{X} + A\vec{V}) - A(A'A)^{-1}A'(\vec{X} + A\vec{V}) \right\|^2 \\ = & \left\| \vec{X} + A\vec{V} - A(A'A)^{-1}A'\vec{X} - A(A'A)^{-1}A'A\vec{V} \right\|^2 \\ = & \left\| \vec{X} + A\vec{V} - A\hat{\vec{\theta}} - A\vec{V} \right\|^2 \\ = & \left\| \vec{X} - A\hat{\vec{\theta}} \right\|^2 \end{aligned}$$

Thus, we may replace $\vec{X}$ be $\vec{X} + A\vec{V}$ and re-estimate $\hat{\vec{\theta}}$ for this new vector without changing the difference between the observed vector and its distance from the predicted vector, $A\hat{\vec{\theta}}$. In particular, we may choose $\vec{V} = -\vec{\theta}$ and set $\vec{Y} = \vec{X} - A\vec{\theta}$, so that $E(\vec{Y}) = E(\vec{X} - A\vec{\theta}) = E(\vec{X}) - A\vec{\theta} = 0$. Since we are adding a fixed number to $\vec{X}$ to give $\vec{Y}$, the covariance matrix does not change, and $Cov(\vec{Y}) = \sigma^2 I$.

Let $\vec{e}_1, ..., \vec{e}_k$ be an orthonormal basis for the column space of A (such a basis exists by the Gram-Schmidt algorithm and the rank of A). Choose $\vec{e}_{k+1}, ..., \vec{e}_n$ such that $\{\vec{e}_1, ..., \vec{e}_n\}$ is an orthonormal basis for R^n. Since $\vec{Y} \in R^n$, we may write $\vec{Y} = \sum_{i=1}^n (\vec{Y}'\vec{e}_j)\vec{e}_j$. This is a linear combination of fixed basis vectors with random coefficients, $\{\vec{Y}'\vec{e}_1, ..., \vec{Y}'\vec{e}_n\}$. The moments of these coefficients are:

$$\begin{aligned} E(\vec{Y}'\vec{e}_j) &= E(\vec{e}_j'\vec{Y}) \\ &= \vec{e}_j' E(\vec{Y}) \\ &= \vec{e}_j'(\vec{0}) \\ &= 0 \end{aligned}$$

$$\begin{aligned}
Cov(\vec{Y}'\vec{e}_j, \vec{Y}'e_h) &= Cov(\vec{e}_j'\vec{Y}, \vec{Y}'e_h) \\
&= E(\vec{e}_j'\vec{Y}\vec{Y}'\vec{e}_h) - E(\vec{e}_j'\vec{Y})E(\vec{Y}'e_h) \\
&= \vec{e}_j'E(\vec{Y}\vec{Y}')\vec{e}_h \\
&= \vec{e}_j'(\sigma^2 I)\vec{e}_h \\
&= \sigma^2(\vec{e}_j'\vec{e}_h) \\
&= \sigma^2 \text{ if } j = h, \text{ 0 otherwise}
\end{aligned}$$

Thus, the coefficients $\{\vec{Y}'\vec{e}_1, ..., \vec{Y}'\vec{e}_n\}$ have mean zero and covariance matrix $\sigma^2 I$. Since $A\hat{\vec{\theta}}$ is the projection of $\vec{Y}$ onto the column space of A, that is, the span of $\{\vec{e}_1, ..., \vec{e}_k\}$, $A\hat{\vec{\theta}} = \sum_{i=1}^{k}(\vec{Y}'\vec{e}_j)\vec{e}_j$, so that $Y - A\hat{\vec{\theta}} = \sum_{i=k+1}^{n}(\vec{Y}'\vec{e}_j)\vec{e}_j$, and $\left\|\vec{Y} - A\hat{\vec{\theta}}\right\|^2 = \sum_{i=k+1}^{n}(\vec{Y}'\vec{e}_j)^2$. Then, $E(\left\|\vec{Y} - A\hat{\vec{\theta}}\right\|^2) = E(\sum_{i=k+1}^{n}(\vec{Y}'\vec{e}_j)^2) = \sum_{i=k+1}^{n} E((\vec{Y}'\vec{e}_j)^2) = \sum_{i=k+1}^{n}\sigma^2 = (n-k)\sigma^2$, since there is zero mean and zero covariance.

21 The joint distribution of the least squares estimates (21)

Suppose $\vec{X}$ is a multivariate normal random vector with mean $A\vec{\theta}$ and covariance matrix $\sigma^2 I$. Let $\hat{\vec{\theta}} = (A'A)^{-1}A'\vec{X}$.

In this case, recall that $\hat{\vec{\theta}}$ is normally distributed with mean $\vec{\theta}$ and covariance matrix $\sigma^2(A'A)^{-1}$.

In the previous theorem, we showed that we may normalize $\vec{X}$ by subtracting its expected value; thus, we assume that $\vec{X}$ has zero mean. Then, we may write $\vec{X} = \sum_{i=1}^{n}(\vec{X}'\vec{e}_j)\vec{e}_j$, $A\hat{\vec{\theta}} = \sum_{i=1}^{k}(\vec{X}'\vec{e}_j)\vec{e}_j$, and $\left\|\vec{X} - A\hat{\vec{\theta}}\right\|^2 = \sum_{i=k+1}^{n}(\vec{X}'\vec{e}_j)^2$, where the coefficients $\{\vec{X}'\vec{e}_1, ..., \vec{X}'\vec{e}_n\}$ have mean zero and covariance matrix $\sigma^2 I$. Then, $\frac{\left\|\vec{X}-A\hat{\vec{\theta}}\right\|^2}{\sigma^2} = \sum_{i=k+1}^{n}(\frac{\vec{X}'\vec{e}_j}{\sigma^2})^2$ is the sum of $n-k$ independent standard normal random variables, which has a chi-squared distribution with $n-k$ degrees of freedom.

Since $\hat{\vec{\theta}} = (A'A)^{-1}A'(\sum_{i=1}^{n}(\vec{X}'\vec{e}_j)\vec{e}_j) = \sum_{i=1}^{n}(\vec{X}'\vec{e}_j)(A'A)^{-1}A'\vec{e}_j$ and $A'\vec{e}_j = 0$ when $j > k$ (since $\vec{e}_j$ is orthogonal to the column space of A in this case), $\hat{\vec{\theta}}$ depends only on $\{(\vec{X}'\vec{e}_1), ..., (\vec{X}'\vec{e}_k)\}$. $\frac{\left\|\vec{X}-A\hat{\vec{\theta}}\right\|^2}{\sigma^2}$ depends only on $\{(\vec{X}'\vec{e}_{k+1}), ..., (\vec{X}'\vec{e}_n)\}$. Since all the $(\vec{X}'\vec{e}_j)$ are independent, these two sets are independent, $\hat{\vec{\theta}}$ and $\frac{\left\|\vec{X}-A\hat{\vec{\theta}}\right\|^2}{\sigma^2}$ are independent.

22 Prediction error (22)

Suppose $\vec{X}$ is a multivariate normal random vector with mean $A\vec{\theta}$ and covariance matrix $\sigma^2 I$. Let $\hat{\vec{\theta}} = (A'A)^{-1}A'\vec{X}$. Let $s^2 = \frac{1}{n-k}\left\|\vec{X} - A\hat{\vec{\theta}}\right\|^2$. Let X_{n+1} be a new observation with associated inputs $\vec{\alpha}$, which is independent of all previous observations. Then, the predicted value of X_{n+1} is $\hat{X}_{n+1} = \vec{\alpha}'\hat{\vec{\theta}}$, and the error of prediction is $\vec{\alpha}'\hat{\vec{\theta}} - X_{n+1} = \vec{\alpha}'(A'A)^{-1}A'\vec{X} - X_{n+1}$. The distribution of the error of prediction is:

$$\begin{aligned} E(\vec{\alpha}'(A'A)^{-1}A'\vec{X} - X_{n+1}) &= E(X_{n+1}) - \vec{\alpha}'(A'A)^{-1}A'E(\vec{X}) \\ &= \vec{\alpha}'\vec{\theta} - \vec{\alpha}'(A'A)^{-1}A'A\vec{\theta} \\ &= \vec{\alpha}'\vec{\theta} - \vec{\alpha}'\vec{\theta} \\ &= 0 \end{aligned}$$

$$\begin{aligned} Var(\vec{\alpha}'\hat{\vec{\theta}} - X_{n+1}) &= Var(\vec{\alpha}'\hat{\vec{\theta}}) + Var(X_{n+1}) \\ &= \vec{\alpha}'(\sigma^2(A'A)^{-1})\vec{\alpha} + \sigma^2 \\ &= \sigma^2(1 + \vec{\alpha}'(A'A)^{-1}\vec{\alpha}) \end{aligned}$$

The estimated standard deviation of the error of prediction is found by replacing σ^2 by s^2, which gives an estimated error of $\sqrt{s^2(1 + \vec{\alpha}'(A'A)^{-1}\vec{\alpha})}$. Notice that X_{n+1} is independent of both s^2 and $\hat{X}_{n+1}$ because they depend only on $X_1, ..., X_n$. In addition, s^2 is independent of $\hat{X}_{n+1}$ because s^2 and $\hat{\vec{\theta}}$ are independent. Thus, we may consider the following ratio which has a t-distribution with $n-k$ degrees of freedom, since the numerator is a standard normal random variable and the denominator is an independent chi-square random variable with $n-k$ degrees of freedom:

$$\frac{\frac{\hat{X}_{n+1} - X_{n+1}}{\sigma\sqrt{1 + \vec{\alpha}'(A'A)^{-1}\vec{\alpha}}}}{\sqrt{\frac{(n-k)s^2/\sigma^2}{n-k}}} = \frac{\hat{X}_{n+1} - X_{n+1}}{s\sqrt{1 + \vec{\alpha}'(A'A)^{-1}\vec{\alpha}}}$$

23 Form of the ANOVA Test (23)

Let $X_{ij}\tilde{}Normal(\mu_i, \sigma^2)$, for $j = 1, ..., n$, $i = 1, ..., k$, with all the X_{ij} independent. Let the null hypothesis be $H_0 : \mu_1 = ... = \mu_k = \mu$. Let the alternative hypothesis be $H_1 : \mu_i \neq \mu_{i'}$ for some $i \neq i'$. We find the likelihood ratio test statistic.

Under the null hypothesis, $X_{ij}\tilde{}Normal(\mu, \sigma^2)$, and this is a sample of size nk from the population $Normal(\mu, \sigma^2)$. The maximum likelihood estimators for μ and σ^2 are then:

$$\begin{aligned}
\hat{\hat{\mu}} &= \frac{1}{nk}\sum_{i=1}^{k}\sum_{j=1}^{n} X_{ij} = \bar{X} \\
\hat{\hat{\sigma}}^2 &= \frac{1}{nk}\sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij} - \bar{X})^2
\end{aligned}$$

Substituting these estimators into the likelihood function gives the restricted maximum likelihood:

$$\begin{aligned}
\max_{\theta \in H_0} L(\vec{X}, \vec{\theta}) &= (2\pi\hat{\hat{\sigma}}^2)^{-nk/2} e^{-\frac{1}{2\hat{\hat{\sigma}}^2}\sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij}-\bar{X})^2} \\
&= (2\pi\hat{\hat{\sigma}}^2)^{-nk/2} e^{-\frac{1}{2\hat{\hat{\sigma}}^2}(nk\hat{\hat{\sigma}}^2)} \\
&= (2\pi\hat{\hat{\sigma}}^2 e)^{nk/2}
\end{aligned}$$

Without this restriction, we instead have k samples of size n with a common variance. Since the means are unrelated, we have

$$\hat{\mu}_i = \frac{1}{n}\sum_{j=1}^{n} X_{ij} = \bar{X}_i$$

We find the maximum likelihood estimator for σ^2 by looking at the log-likelihood:

$$\begin{aligned}
\log L(\mu_1, ..., \mu_n, \sigma^2) &= \log((2\pi\sigma^2)^{-nk/2} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij}-\mu_i)^2} \\
&= -\frac{nk}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij} - \mu_i)^2
\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial\sigma^2}\log L(\mu_1, ..., \mu_n, \sigma^2) &= -\frac{nk}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij} - \mu_i)^2 \\
&= (\sigma^2 - \frac{1}{nk}\sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij} - \mu_i)^2)(-\frac{nk}{2(\sigma^2)^2})
\end{aligned}$$

Solving for σ^2 and substituting the maximum likelihood estimators for the μ_i, we find:

$$\hat{\sigma}^2 = \frac{1}{nk}\sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij} - \bar{X}_i)^2$$

Substituting these estimators gives the unrestricted maximum likelihood:

$$\begin{aligned}
\max_{\theta \in H} L(\vec{X}, \vec{\theta}) &= (2\pi\hat{\sigma}^2)^{-nk./2} e^{-\frac{1}{2\hat{\sigma}^2}(\sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij}-\bar{X}_i)^2)} \\
&= (2\pi\hat{\sigma}^2)^{-nk./2} e^{-\frac{1}{2\hat{\sigma}^2}(nk\hat{\sigma}^2)} \\
&= (2\pi\hat{\sigma}^2 e)^{-nk/2}
\end{aligned}$$

We find the likelihood ratio test statistic by looking at the ratio of the restricted and unrestricted maximum likelihoods:

$$\lambda(\vec{X}) = \frac{(2\pi\hat{\sigma}^2 e)^{nk/2}}{(2\pi\hat{\hat{\sigma}}^2 e)^{-nk/2}} = (\frac{\hat{\sigma}^2}{\hat{\hat{\sigma}}^2})^{nk/2} = (\frac{\sum_{i=1}^k \sum_{j=1}^n (X_{ij} - \bar{X}_i)^2}{\sum_{i=1}^k \sum_{j=1}^n (X_{ij} - \bar{X})^2})^{nk/2}$$

Notice that

$$\begin{aligned}
\sum_{i=1}^k \sum_{j=1}^n (X_{ij} - \bar{X})^2 &= \sum_{i=1}^k \sum_{j=1}^n (X_{ij} - \bar{X}_i + \bar{X}_i - \bar{X})^2 \\
&= \sum_{i=1}^k (\sum_{j=1}^n (X_{ij} - \bar{X}_i)^2 + n(\bar{X}_i - \bar{X})^2) \\
&= \sum_{i=1}^k \sum_{j=1}^n (X_{ij} - \bar{X}_i)^2 + n\sum_{i=1}^k (\bar{X}_i - \bar{X})^2
\end{aligned}$$

The first term in this equation is defined as the sum of squares within (SSW); the second term is defined as the sum of squares between (SSB). Using these definition, we may rewrite the likelihood ratio as:

$$\begin{aligned}
\lambda(\vec{X}) &= (\frac{\sum_{i=1}^k \sum_{j=1}^n (X_{ij} - \bar{X}_i)^2}{\sum_{i=1}^k \sum_{j=1}^n (X_{ij} - \bar{X})^2})^{nk/2} \\
&= (\frac{SSW}{SSW + SSB})^{nk/2} \\
&= (\frac{1}{1 + \frac{SSB}{SSW}})^{nk/2}
\end{aligned}$$

Thus, we see that $\lambda(\vec{X})$ is a decreasing function of $\frac{SSB}{SSW}$.

24 The ANOVA test statistic has an F-distribution (24)

Recall that $SSB = n\sum_{i=1}^k (\bar{X}_i - \bar{X})^2$. Since we may write $\bar{X} = \frac{1}{k}\sum_{i=1}^k (\frac{1}{n}\sum_{i=1}^n X_{ij}) = \frac{1}{k}\sum_{i=1}^k \bar{X}_i$, SSB can be written as a function of only $\bar{X}_1, ..., \bar{X}_k$. Since $\bar{X}_1, ..., \bar{X}_k$ are the means of disjoint independent samples of size n from $Normal(\mu, \sigma^2)$, $\bar{X}_1, ..., \bar{X}_k$ are independent and distributed $Normal(\mu, \frac{\sigma^2}{n})$. Thus, $\sum_{i=1}^k (\bar{X}_i - \bar{X})^2$ is the sum of squares about the average of a sample of size k from a distribution with variance $\frac{\sigma^2}{n}$, which means that $\frac{\sum_{i=1}^k (\bar{X}_i - \bar{X})^2}{\sigma^2/n} = \frac{SSB}{\sigma^2}$ has a chi-squared distribution with $k-1$ degrees of freedom.

Recall that $SSW = \sum_{i=1}^k \sum_{j=1}^n (X_{ij} - \bar{X}_i)^2$. Let $s_i^2 = \frac{1}{n-1}\sum_{j=1}^n (X_{ij} - \bar{X}_i)^2$

for each $i = 1, ...k$. Then we may rewrite

$$\begin{aligned} SSW &= \sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij} - \bar{X}_i)^2 \\ &= \sum_{i=1}^{k}(n-1)s_i^2 \end{aligned}$$

Because $s_1^2, ..., s_k^2$ are the sample variances of samples of size n from normal populations with variance σ^2, $\frac{(n-1)s_1^2}{\sigma^2}, ..., \frac{(n-1)s_k^2}{\sigma^2}$ are each distributed chi-squared with $n-1$ degrees of freedom. Since $s_1^2, ..., s_k^2$ depend on disjoint independent samples, they are independent. Thus, $\frac{\sum_{i=1}^{k}(n-1)s_i^2}{\sigma^2} = \frac{SSW}{\sigma^2}$ is the sum of k independent chi-squared variables with $n-1$ degrees of freedom. By the Addition Theorem for Chi-Squares, this sum has a chi-squared distribution with $k(n-1)$ degrees of freedom.

Consider $\bar{X}_i$ and $s_{i'}^2$. If $i = i'$, then these are the mean and sample variance from a sample from a normal population. These two statistics are independent. If $i \neq i'$, then these two statistics depend on different independent samples, and are therefore independent. Thus, $\{\bar{X}_1, ..., \bar{X}_k\}$ and $\{s_1^2, ..., s_k^2\}$ are independent sets. Since SSB and SSW depend on these two sets, SSB and SSW are independent. Therefore, the following expression is the ratio of independent chi-square random variables with $k-1$ and $k(n-1)$ degrees of freedom respectively, divided by their degrees of freedom:

$$F = \frac{\frac{SSB}{\sigma^2}/(k-1)}{\frac{SSW}{\sigma^2}/k(n-1)} = \frac{SSB/(k-1)}{SSW/k(n-1)}$$

Hence, the test statistic for ANOVA, $\frac{SSB/(k-1)}{SSW/k(n-1)}$ has an F-distribution with $k-1$ and $k(n-1)$ degrees of freedom.

25 Multivariate Central Limit Theorem (25)

Theorem 3 *Let $\vec{X}_1, ..., \vec{X}_n$ be a sample of independent identically distributed random vectors in R^m. Let $E(\vec{X}_i) = \vec{\mu}$ and $Cov(\vec{X}_i) = \Sigma$. Then, as $n \to \infty$, $\frac{1}{\sqrt{n}}\sum_{j=1}^{n}(\vec{X}_j - \vec{\mu})$ has the limiting distribution $Normal_m(\vec{0}, \Sigma)$.*

Proof. Let $\vec{t} \in R^m$ be any fixed vector. Define $\xi_n = \vec{t}'(\frac{1}{\sqrt{n}}\sum_{j=1}^{n}(\vec{X}_j - \vec{\mu})) = \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\vec{t}'(\vec{X}_j - \vec{\mu})$. Each $\vec{t}'(\vec{X}_j - \vec{\mu})$ is an independent, identically distributed random variable with:

$$\begin{aligned} E(\vec{t}'(\vec{X}_j - \vec{\mu})) &= \vec{t}'E(\vec{X}_j - \vec{\mu}) \\ &= \vec{t}'E(\vec{X}_j) - \vec{t}'\vec{\mu} \\ &= \vec{t}'\vec{\mu} - \vec{t}'\vec{\mu} \\ &= 0 \end{aligned}$$

$$\begin{aligned} Var(\vec{t}'(\vec{X}_j - \vec{\mu})) &= \vec{t}'Cov(\vec{X}_j - \vec{\mu})\vec{t} \\ &= \vec{t}'Cov(\vec{X}_j)\vec{t} \\ &= \vec{t}'\Sigma\vec{t} \end{aligned}$$

By the one-variable central limit theorem, $\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\vec{t}'(\vec{X}_j - \vec{\mu})$ has a limiting distribution with mean 0 and variance $\vec{t}'\Sigma\vec{t}$.

By a theorem on the convergence of characteristic functions, the convergence of this distribution to a normal distribution implies the convergence of its characteristic function to the characteristic function of the same normal distribution. That is,

$$\lim_{n\to\infty} E(e^{i\frac{u}{\sqrt{n}}\sum_{j=1}^{n}\vec{t}'(\vec{X}_j-\vec{\mu})}) = e^{-\frac{1}{2}u^2\vec{t}'\Sigma\vec{t}}$$

for all $u \in R$. This is true for all $\vec{t} \in R^m$. In particular, set $u = 1$. Then, we find

$$\lim_{n\to\infty} E(e^{i\vec{t}'(\frac{1}{\sqrt{n}}\sum_{j=1}^{n}(\vec{X}_j-\vec{\mu}))}) = e^{-\frac{1}{2}\vec{t}'\Sigma\vec{t}}$$

for all $\vec{t} \in R^m$, and every linear combination of the elements of $\frac{1}{\sqrt{n}}\sum_{j=1}^{n}(\vec{X}_j - \vec{\mu})$ converges to a normal distribution. By the Multivariate Continuity Theorem, the convergence of every linear combination elements of a random vector to a normal distribution implies the convergence of the random vector to a multivariate normal distribution. Thus, $\frac{1}{\sqrt{n}}\sum_{j=1}^{n}(\vec{X}_j - \vec{\mu})$ converges to a $Normal(\vec{0}, \Sigma)$ distribution. ∎

26 Random vectors describing multinomial distributions (26)

Let $\vec{Z} = (Z_1, ..., Z_k)'$ be a random vector with $P(Z_i = 1) = p_i$, $\sum_{i=1}^{k} Z_i = \sum_{i=1}^{k} p_i = 1$. Because each Z_i is a binomial random variable with probability p_i of success, $E(Z_i) = p_i$ and $Var(Z_i) = p_i(1 - p_i)$. Because exactly one of $Z_1, ..., Z_k$ is one, $Z_iZ_j = 0$ when $i \neq j$. Therefore, we find that $Cov(Z_i, Z_j) = E(Z_iZ_j) - E(Z_i)E(Z_j) = 0 - p_ip_j = -p_ip_j$. This means that the covariance matrix of $\vec{Z}$ is:

$$\begin{aligned} Cov(\vec{Z}) &= \begin{bmatrix} p_1(1-p_1) & ... & -p_ip_j \\ ... & ... & ... \\ -p_ip_j & ... & p_k(1-p_k) \end{bmatrix} \\ &= \begin{bmatrix} p_1 & ... & 0 \\ .. & ... & ... \\ 0 & ... & p_k \end{bmatrix} - \begin{bmatrix} & ... & \\ ... & -p_ip_j & ... \\ & ... & \end{bmatrix} \\ &= \begin{bmatrix} p_1 & ... & 0 \\ .. & ... & ... \\ 0 & ... & p_k \end{bmatrix} - \begin{bmatrix} p_1 \\ ... \\ p_k \end{bmatrix} \begin{bmatrix} p_1 & ... & p_k \end{bmatrix} \end{aligned}$$

Let P be the first matrix in this expansion; P is a diagonal matrix with the i^{th} diagonal entry equal to p_i. Then, $P^{-\frac{1}{2}}$ is a diagonal matrix with entries $\frac{1}{\sqrt{p_i}}$ along the diagonal, and:

$$\begin{aligned} Cov(P^{-\frac{1}{2}}\vec{Z}) &= P^{-\frac{1}{2}}Cov(\vec{Z})P^{-\frac{1}{2}} \\ &= P^{-\frac{1}{2}}(P - \begin{bmatrix} p_1 \\ ... \\ p_k \end{bmatrix} \begin{bmatrix} p_1 & ... & p_k \end{bmatrix})P^{-\frac{1}{2}} \\ &= I - \begin{bmatrix} \sqrt{p_1} \\ ... \\ \sqrt{p_k} \end{bmatrix} \begin{bmatrix} \sqrt{p_1} & ... & \sqrt{p_k} \end{bmatrix} \end{aligned}$$

27 The limiting distribution of the chi-square test statistic (27)

Consider a sample of size n in which each observation falls into exactly one of k classes, with probability p_i of being in class i, $\sum_{i=1}^{k} p_i = 1$. Let f_i be the sample frequency of class i, so that $\sum_{i=1}^{k} f_i = n$. Define $\Xi = \sum_{i=1}^{k}(\frac{f_i - np_i}{\sqrt{np_i}})^2$.

Let P be the diagonal matrix with diagonal entries p_i. Let $\vec{Z}_j$ be a vector with i^{th} entry equal to 1 if the j^{th} observation falls into the i^{th} category and 0 otherwise. Then, $\sum_{j=1}^{n} \vec{Z}_j = \begin{bmatrix} f_1 \\ ... \\ f_k \end{bmatrix}$ and $E(\sum_{j=1}^{n} \vec{Z}_j) = \sum_{j=1}^{n} E(\vec{Z}_j) =$

$\sum_{j=1}^{n} \begin{bmatrix} p_1 \\ ... \\ p_k \end{bmatrix} = \begin{bmatrix} np_1 \\ ... \\ np_k \end{bmatrix}$. Then, we find:

$$\begin{aligned}
\Xi &= \sum_{i=1}^{k} (\frac{f_i - np_i}{\sqrt{np_i}})^2 \\
&= \left\| \begin{bmatrix} \frac{f_1 - np_1}{\sqrt{np_1}} \\ ... \\ \frac{f_k - np_k}{\sqrt{np_k}} \end{bmatrix} \right\|^2 \\
&= \left\| \frac{1}{\sqrt{n}} \begin{bmatrix} \frac{f_1 - np_1}{\sqrt{p_1}} \\ ... \\ \frac{f_k - np_k}{\sqrt{p_k}} \end{bmatrix} \right\|^2 \\
&= \left\| \frac{1}{\sqrt{n}} P^{-\frac{1}{2}} \begin{bmatrix} f_1 - np_1 \\ ... \\ f_k - np_k \end{bmatrix} \right\|^2 \\
&= \left\| \frac{1}{\sqrt{n}} P^{-\frac{1}{2}} (\begin{bmatrix} f_1 \\ ... \\ f_k \end{bmatrix} - \begin{bmatrix} np_1 \\ ... \\ np_k \end{bmatrix}) \right\|^2 \\
&= \left\| \frac{1}{\sqrt{n}} P^{-\frac{1}{2}} (\sum_{j=1}^{n} \vec{Z}_j - \sum_{j=1}^{n} E(\vec{Z}_j)) \right\|^2 \\
&= \left\| \frac{1}{\sqrt{n}} P^{-\frac{1}{2}} \sum_{j=1}^{n} (\vec{Z}_j - E(\vec{Z}_j)) \right\|^2
\end{aligned}$$

By the multivariate central limit theorem, $\frac{1}{\sqrt{n}} \sum_{j=1}^{n} (\vec{Z}_j - E(\vec{Z}_j))$ has a limiting normal distribution with mean $\vec{0}$ and covariance matrix $Cov(\vec{Z})$. Thus, $\frac{1}{\sqrt{n}} P^{-\frac{1}{2}} \sum_{j=1}^{n} (\vec{Z}_j - E(\vec{Z}_j))$ has the limiting distribution $Normal(\vec{0}, P^{-\frac{1}{2}} Cov(\vec{Z}) P^{-\frac{1}{2}}) = Normal(\vec{0}, I - \begin{bmatrix} \sqrt{p_1} \\ ... \\ \sqrt{p_k} \end{bmatrix} \begin{bmatrix} \sqrt{p_1} & ... & \sqrt{p_k} \end{bmatrix})$. Let $\vec{Y} = \frac{1}{\sqrt{n}} P^{-\frac{1}{2}} \sum_{j=1}^{n} (\vec{Z}_j - E(\vec{Z}_j))$.

Let Q be an orthogonal matrix with first row $\begin{bmatrix} \sqrt{p_1} & ... & \sqrt{p_k} \end{bmatrix}$. Since the norm is invariant under orthogonal transformations, $\left\| \vec{Y} \right\|^2 = \left\| Q\vec{Y} \right\|^2$. Also, $Q\vec{Y}$ is distributed approximately $Normal(\vec{0}, QCov(\vec{Y})Q')$, so that the covariance matrix is:

$$\begin{aligned}
QCov(\vec{Y})Q' &= Q(I - \begin{bmatrix}\sqrt{p_1}\\ ...\\ \sqrt{p_k}\end{bmatrix}\begin{bmatrix}\sqrt{p_1} & ... & \sqrt{p_k}\end{bmatrix})Q' \\
&= QQ' - (Q\begin{bmatrix}\sqrt{p_1}\\ ...\\ \sqrt{p_k}\end{bmatrix})(Q\begin{bmatrix}\sqrt{p_1}\\ ...\\ \sqrt{p_k}\end{bmatrix})' \\
&= I - \begin{bmatrix}1\\ 0\\ ...\\ 0\end{bmatrix}\begin{bmatrix}1 & 0 & ... & 0\end{bmatrix} \\
&= \begin{bmatrix}0 & \vec{0}\\ \vec{0} & I_{k-1}\end{bmatrix}
\end{aligned}$$

Thus, we see that $Q\vec{Y}$ can be written $\begin{bmatrix}0\\ X_2\\ ...\\ X_k\end{bmatrix}$; where the X_i are independent, approximately standard normal variables. Thus, $\left\|\vec{Y}\right\|^2 = \left\|Q\vec{Y}\right\|^2 = \sum_{i=2}^{k} X_i^2$ is (approximately) the sum of $k-1$ independent standard normal random variables and therefore is (approximately) χ^2 with $k-1$ degrees of freedom. Thus, the chi-square statistic has a limiting χ^2 distribution with $k-1$ degrees of freedom.

28 The limiting normal distribution of quantile statistics. (28)

Let $F(x)$ be a cumulative density function. Let $F'(x)$ exist. Let ξ_p be the p^{th} quantile of $F(x)$ (that is, $F(\xi_p) = p$). Assume $F'(\xi_p) > 0$. Let $r = r(n)$ be an index of an order statistic, $X_{(r)}$, such that $\lim_{n\to\infty} \sqrt{n}(\frac{r(n)}{n} - p) = 0$. We will show that $\sqrt{n}(X_{(r)} - \xi_p)$ has a limiting distribution which is normal with mean 0 and variance $\frac{p(1-p)}{(F'(\xi_p))^2}$.

Consider the cumulative distribution of $\sqrt{n}(X_{(r)} - \xi_p)$:

$$\begin{aligned}
P(\sqrt{n}(X_{(r)} - \xi_p) &\leq y) = P(X_{(r)} \leq \frac{y}{\sqrt{n}} + \xi_p) \\
&= \sum_{j=r}^{n} \binom{n}{j} (F(\frac{y}{\sqrt{n}} + \xi_p))^j (1 - F(\frac{y}{\sqrt{n}} + \xi_p))^{n-j} \\
&= P(\text{at least r successes of n} \mid \text{probability } F(\frac{y}{\sqrt{n}} + \xi_p) \text{ of success}) \\
&= 1 - P(\text{less than r successes})
\end{aligned}$$

According to the normal approximation to the binomial distribution, a binomial distribution with n trials and probability θ of success is approximately normal with mean $n\theta$ and variance $\frac{1}{n\theta(1-\theta)}$. In this case, $\theta = F(\frac{y}{\sqrt{n}} + \xi_p)$ and we find:

$$\lim_{n\to\infty} P(\sqrt{n}(X_{(r)} - \xi_p) \leq y) = 1 - \Phi(\lim_{n\to\infty} \frac{r - nF(\frac{y}{\sqrt{n}} + \xi_p)}{\sqrt{nF(\frac{y}{\sqrt{n}} + \xi_p)(1 - F(\frac{y}{\sqrt{n}} + \xi_p))}})$$

However, we must show that $\lim_{n\to\infty} \frac{r-nF(\frac{y}{\sqrt{n}}+\xi_p)}{\sqrt{nF(\frac{y}{\sqrt{n}}+\xi_p)(1-F(\frac{y}{\sqrt{n}}+\xi_p))}}$ exists.

Case 1: Uniform distribution.
In the case of the uniform distribution:

$$\begin{aligned} F(x) &= x, x \in [0,1] \\ \xi_p &= p \\ F(\xi_p + \frac{y}{\sqrt{n}}) &= \xi_p + \frac{y}{\sqrt{n}} = p + \frac{y}{\sqrt{n}} \\ F'(\xi_p) &= 1 \end{aligned}$$

Substituting these values into the limit, we find:

$$\begin{aligned} \lim_{n\to\infty} \frac{r - nF(\frac{y}{\sqrt{n}} + \xi_p)}{\sqrt{nF(\frac{y}{\sqrt{n}} + \xi_p)(1 - F(\frac{y}{\sqrt{n}} + \xi_p))}} &= \lim_{n\to\infty} \frac{r - n(p + \frac{y}{\sqrt{n}})}{\sqrt{n(p + \frac{y}{\sqrt{n}})(1 - p - \frac{y}{\sqrt{n}})}} \\ &= \lim_{n\to\infty} \frac{r - np - y\sqrt{n}}{\sqrt{np(1-p)}} \\ &= \frac{1}{\sqrt{p(1-p)}} \lim_{n\to\infty} (\frac{r - np}{\sqrt{n}} - y) \\ &= \frac{1}{\sqrt{p(1-p)}} \lim_{n\to\infty} (\sqrt{n}(\frac{r}{n} - p) - y) \\ &= -\frac{y}{\sqrt{p(1-p)}} \end{aligned}$$

(Note: $\sqrt{p + \frac{y}{\sqrt{n}}}$ converges to $\sqrt{p}$ faster than $p + \frac{y}{\sqrt{n}}$ converges to p.)
Applying the normal approximation, we find:

$$\begin{aligned} P(\sqrt{n}(X_{(r)} - \xi_p) &\leq y) = 1 - \Phi(-\frac{y}{\sqrt{p(1-p)}}) \\ &= \Phi(\frac{y}{\sqrt{p(1-p)}}) \end{aligned}$$

Case 2: The general case.

By the Probability Integral Transformation, if $F(x)$ is a continuous cumulative distribution function and U is a random variable uniformly distributed on $[0,1]$, then the cumulative distribution function of the random variable $F(U)$ is $F(x)$. Conversely, if X is a random variable with cumulative distribution function $F(x)$, then the random variable $F(X)$ is randomly distributed uniform on $[0,1]$.

Since $F(x)$ is a non-decreasing function (and is increasing in a neighborhood of ξ_p),

$$P(X_{(r)} \leq \xi_p + \frac{y}{\sqrt{n}}) = P(F(X_{(r)}) \leq F(\xi_p + \frac{y}{\sqrt{n}}))$$

Applying a Taylor expansion about ξ_p, we find:

$$\begin{aligned} P(F(X_{(r)}) &\leq F(\xi_p + \frac{y}{\sqrt{n}})) = P(F(X_{(r)}) \leq F(\xi_p) + F'(\xi_p)\frac{y}{\sqrt{n}} + \varepsilon(y)) \\ &= P(\sqrt{n}(F(X_{(r)}) - F(\xi_p)) \leq F'(\xi_p)y + \varepsilon(y)\sqrt{n}) \end{aligned}$$

($\varepsilon(y)$ is a smaller order function containing the rest of the terms in the Taylor expansion; it is negligible.) Notice that $F(X_{(r)}) = U_{(r)}$, where $U_{(r)}$ is the r^{th} order statistic of a sample of n uniform random variables on $[0,1]$. Also, recall that $F(\xi_p) = p$. Substituting these facts, applying Case 1, and taking the limit, we find:

$$\begin{aligned} P(X_{(r)} &\leq \xi_p + \frac{y}{\sqrt{n}}) = P(\sqrt{n}(U_{(r)} - p) \leq F'(\xi_p)y + \varepsilon(y)\sqrt{n}) \\ &= \Phi(\frac{1}{\sqrt{p(1-p)}}F'(\xi_p)y) \end{aligned}$$

Thus,

$$\begin{aligned} P(\sqrt{n}(X_{(r)} - \xi_p) &\leq y) = P(X_{(r)} \leq \xi_p + \frac{y}{\sqrt{n}}) \\ &= \Phi(\frac{1}{\sqrt{p(1-p)}}F'(\xi_p)y) \end{aligned}$$

which is equivalent to $\sqrt{n}(X_{(r)} - \xi_p)\tilde{}Normal(0, \frac{p(1-p)}{(F'(\xi_p))^2})$.

29 Propogation of Errors (29)

Theorem 4 *Let $\{X_n\}$ be a sequence of random variables. Let $\{a_n\}$ be a sequence of constants such that $a_n > 0$ for all n and $\lim_{n\to\infty} a_n = 0$. Let μ be fixed. If $\frac{X_n-\mu}{a_n}$ has a limiting $Normal(0,\sigma^2)$ distribution, then for any continuously differentiable function $f : R \to R$, $\frac{f(X_n)-f(\mu)}{a_n}$ has a limiting $Normal(0,\sigma^2(f'(\mu))^2)$ distribution.*

Proof. For every $\varepsilon > 0$, $P(|X_n - \mu| > \varepsilon) = P(\left|\frac{X_n-\mu}{a_n}\right| > \frac{\varepsilon}{a_n})$. Since ε is fixed, $\lim_{n\to\infty} \frac{\varepsilon}{a_n} = \infty$, so that $\lim_{n\to\infty} P(|X_n - \mu| > \varepsilon) = \lim_{n\to\infty} P(\left|\frac{X_n-\mu}{a_n}\right| > \frac{\varepsilon}{a_n}) = 0$, since $\frac{\varepsilon}{a_n}$ diverges while X_n and μ are finite. Thus, X_n converges in probability to μ. By the Taylor expansion, $(f(X_n) - f(\mu)) \approx f'(\mu)(X_n - \mu)$ within ε of μ. Since $\frac{X_n-\mu}{a_n}$ has a limiting $Normal(0, \sigma^2)$ distribution, $f'(\mu)\frac{X_n-\mu}{a_n} = \frac{f(X_n)-f(\mu)}{a_n}$ has a normal limiting distribution with mean $f'(\mu)0 = 0$ and variance $(f'(\mu))^2\sigma^2$. ■

Theorem 5 *Let $\{\vec{X}_n\}$ be a sequence of random vectors. Let $\{a_n\}$ be a sequence of constants such that $a_n > 0$ for all n and $\lim_{n\to\infty} a_n = 0$. Let $\vec{\mu}$ be a fixed vector. If $\frac{1}{a_n}(\vec{X}_n - \vec{\mu})$ has a limiting $Normal(\vec{0}, \Sigma)$ distribution, then for any smooth function $f : R^k \to R$, $\frac{1}{a_n}(f(\vec{X}_n) - f(\vec{\mu}))$ has the limiting distribution $Normal(\vec{0}, ((\nabla f)(\mu))'\Sigma((\nabla f)(\mu)))$*

Proof. Because $\frac{1}{a_n}(\vec{X}_n - \vec{\mu})$ has a limiting distribution, $\vec{X}_n$ converges in probability to $\vec{\mu}$. By the vector form of Taylor"s expansion, $f(\vec{X}_n) - f(\vec{\mu}) \approx ((\nabla f)(\mu))'(\vec{X}_n - \mu)$. Since $\frac{1}{a_n}(\vec{X}_n - \vec{\mu})$ has a limiting $Normal(\vec{0}, \Sigma)$ distribution, $\frac{1}{a_n}((\nabla f)(\mu))'(\vec{X}_n - \mu) = \frac{1}{a_n}(f(\vec{X}_n) - f(\vec{\mu}))$ has a limiting normal distribution with mean $((\nabla f)(\mu))'\vec{0} = 0$ and variance $((\nabla f)(\mu))'\Sigma((\nabla f)(\mu))$. ■

30 Distribution of the sample correlation coefficient (30)

Definition 6 *Let X and Y be bivariate normal random variables with correlation coefficient ρ. Let $(X_1, Y_1), ..., (X_n, Y_n)$ be a sample of pairs from this distribution. Let $r = \frac{\sum_{i=1}^n (X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{(\sum_{i=1}^n (X_i-\bar{X})^2)(\sum_{i=1}^n (Y_i-\bar{Y})^2)}}$. We call r the sample correlation coefficient.*

Lemma 7 *ρ and r are invariant under positive linear transformations, $U = aX + b$ and $V = cY + d$, with $a > 0, c > 0$.*

Proof. Let $\rho(U, V)$ be the correlation coefficient of U and V and $\rho(X, Y)$ be the correlation coefficient of X and Y. Then,

$$\rho(X, Y) = E(\frac{X - E(X)}{\sqrt{Var(X)}} \cdot \frac{Y - E(Y)}{\sqrt{Var(Y)}})$$

$$E(U) = aE(X) + b$$

$$E(V) = cE(Y) + d$$

$$Var(U) = a^2 Var(X)$$

$$Var(V) = c^2 Var(Y)$$

$$
\begin{aligned}
\rho(U,V) &= E(\frac{U-E(U)}{\sqrt{Var(U)}} \cdot \frac{V-E(V)}{\sqrt{Var(V)}}) \\
&= E(\frac{aX+b-(aE(X)+b)}{\sqrt{a^2Var(X)}} \cdot \frac{cY+d-(cE(Y)+d)}{\sqrt{c^2Var(Y)}}) \\
&= E(\frac{X-E(X)}{\sqrt{Var(X)}} \cdot \frac{Y-E(Y)}{\sqrt{Var(Y)}}) \\
&= \rho(X,Y)
\end{aligned}
$$

We do the same for $r(X,Y)$ and $r(U,V)$:

$$r(X,Y) = \frac{\sum_{i=1}^n (X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{(\sum_{i=1}^n (X_i-\bar{X})^2)(\sum_{i=1}^n (Y_i-\bar{Y})^2)}}$$

$$\bar{U} = a\bar{X}+b$$

$$\bar{V} = c\bar{Y}+d$$

$$\sum_{i=1}^n (U_i-\bar{U})^2 = \sum_{i=1}^n (aX_i+b-(a\bar{X}+b))^2 = \sum_{i=1}^n a^2(X_i-\bar{X})^2$$

$$\sum_{i=1}^n (V_i-\bar{V})^2 = \sum_{i=1}^n (cY_i+d-(c\bar{Y}+d))^2 = \sum_{i=1}^n c^2(Y_i-\bar{Y})^2$$

$$
\begin{aligned}
\sum_{i=1}^n (U_i-\bar{U})(V_i-\bar{V}) &= \sum_{i=1}^n (aX_i+b-(a\bar{X}+b))(cY_i+d-(c\bar{Y}+d)) \\
&= \sum_{i=1}^n ac(X_i-\bar{X})(Y_i-\bar{Y})
\end{aligned}
$$

$$
\begin{aligned}
r(U,V) &= \frac{\sum_{i=1}^n (U_i-\bar{U})(V_i-\bar{V})}{\sqrt{(\sum_{i=1}^n (U_i-\bar{U})^2)(\sum_{i=1}^n (V_i-\bar{V})^2)}} \\
&= \frac{\sum_{i=1}^n ac(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{(\sum_{i=1}^n a^2(X_i-\bar{X})^2)(\sum_{i=1}^n c^2(Y_i-\bar{Y})^2)}} \\
&= \frac{\sum_{i=1}^n (X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{(\sum_{i=1}^n (X_i-\bar{X})^2)(\sum_{i=1}^n (Y_i-\bar{Y})^2)}} \\
&= r(X,Y)
\end{aligned}
$$

■

Theorem 8 *As $n \to \infty$, $\sqrt{n}(r-\rho)$ has a limiting $Normal(0,(1-\rho^2)^2)$ distribution.*

Proof.

Lemma 9 ***Proof.*** *Because ρ and the distribution of r are invariant under positive linear transformations, given any bivariate normal random variables, we may subtract their means and divide by their standard deviations without affect ρ or r. Thus, without loss of generality, we may prove this theorem for standard bivariate normal random variables, X and Y, with correlation coefficient, ρ.* ■

Lemma 10 *For standard bivariate normals, r has the same sampling distribution as $\frac{\sum_{i=2}^{n} V_i W_i}{\sqrt{(\sum_{i=2}^{n} V_i^2)(\sum_{i=2}^{n} W_i^2)}}$, where $\vec{V}$ and $\vec{W}$ are standard bivariate normals with the same correlation coefficient.*

Proof. Let Q be the orthonormal transformation with first row $\begin{bmatrix} \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \end{bmatrix}$. Then, the first elements of $\vec{V} = Q\vec{X}$ and $\vec{W} = Q\vec{Y}$ are $\sqrt{n}\bar{X}$ and $\sqrt{n}\bar{Y}$ respectively. Because orthonormal transformations preserve inner products,

$$\begin{aligned} \sum_{i=1}^{n}(X_i - \bar{X})^2 &= \sum_{i=1}^{n} X_i^2 - n\bar{X}^2 \\ &= \left\|\vec{V}\right\|^2 - V_1^2 \\ &= \sum_{i=1}^{n} V_i^2 \end{aligned}$$

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n} W_i^2$$

$$\begin{aligned} \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) &= \sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y} \\ &= < \vec{V}, \vec{W} > - V_1 W_1 \\ &= \sum_{i=2}^{n} V_i W_i \end{aligned}$$

Because Q is an orthogonal transformation, it is a rotation of R^n. Therefore, $\vec{V}$ and $\vec{W}$ have the same correlation as $\vec{X}$ and $\vec{Y}$. ■

Define $f(u_1, u_2, u_3) = \frac{u_3}{\sqrt{u_1 u_2}}$. Then, $r = f(\sum_{i=1}^{n}(X_i - \bar{X})^2, \sum_{i=1}^{n}(Y_i - \bar{Y})^2, \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})) = f(\sum_{i=1}^{n} V_i^2, \sum_{i=1}^{n} W_i^2, \sum_{i=2}^{n} V_i W_i)$. Define independent random vectors $\vec{Z}_i = \begin{bmatrix} V_1^2 \\ W_i^2 \\ V_i W_i \end{bmatrix}$. Since V_i and W_i are standard bivariate normal, $E(V_i^2) = E(W_i^2) = 1$ and $E(V_i W_i) = \rho$. Thus, $E(\vec{Z}_i) = \begin{bmatrix} 1 \\ 1 \\ \rho \end{bmatrix}$.

It can be shown that $Cov(\vec{Z}_i) = 2\begin{bmatrix}1 & \rho^2 & \rho \\ \rho^2 & 1 & \rho \\ \rho & \rho & \frac{1+\rho^2}{2}\end{bmatrix}$. Thus, by the Multivariate Central Limit Theorem, $\frac{1}{\sqrt{n-1}}\sum_{i=2}^n(\begin{bmatrix}V_1^2 \\ W_i^2 \\ V_iW_i\end{bmatrix} - \begin{bmatrix}1 \\ 1 \\ \rho\end{bmatrix})$ has a limiting $Normal(\vec{0}, 2\begin{bmatrix}1 & \rho^2 & \rho \\ \rho^2 & 1 & \rho \\ \rho & \rho & \frac{1+\rho^2}{2}\end{bmatrix})$ distribution. By the Propogation of Error Theorem, $\sqrt{n-1}(f(\frac{1}{n-1}\sum_{i=2}^n V_i^2, \frac{1}{n-1}\sum_{i=2}^n W_i^2, \frac{1}{n-1}\sum_{i=2}^n V_iW_i) - f(1,1,\rho))$ has a limiting $Normal(\vec{0}, ((\nabla f)(\mu))' Cov(\vec{Z}_i)((\nabla f)(\mu))$ distribution. We calculate:

$$\nabla f(u_1,u_2,u_3) = \begin{bmatrix}\frac{u_3}{\sqrt{u_2}}(-\frac{1}{2})(u_1)^{-\frac{3}{2}} \\ \frac{u_3}{\sqrt{u_1}}(-\frac{1}{2})(u_2)^{-\frac{3}{2}} \\ \frac{1}{\sqrt{u_1u_2}}\end{bmatrix}$$

$$(\nabla f)(1,1,\rho) = \begin{bmatrix}-\frac{1}{2}\rho \\ -\frac{1}{2}\rho \\ 1\end{bmatrix}$$

$$\begin{aligned}((\nabla f)(1,1,\rho))' Cov(\vec{Z}_i)((\nabla f)(1,1,\rho)) &= 2\begin{bmatrix}-\frac{1}{2}\rho & -\frac{1}{2}\rho & 1\end{bmatrix}\begin{bmatrix}1 & \rho^2 & \rho \\ \rho^2 & 1 & \rho \\ \rho & \rho & \frac{1+\rho^2}{2}\end{bmatrix}\begin{bmatrix}-\frac{1}{2}\rho \\ -\frac{1}{2}\rho \\ 1\end{bmatrix} \\ &= 2\begin{bmatrix}-\frac{1}{2}\rho & -\frac{1}{2}\rho & 1\end{bmatrix}\begin{bmatrix}-\frac{1}{2}\rho^3 + \frac{1}{2}\rho \\ -\frac{1}{2}\rho^3 + \frac{1}{2}\rho \\ \frac{1}{2} - \frac{1}{2}\rho^2\end{bmatrix} \\ &= \rho^4 - 2\rho^2 + 1 \\ &= (1-\rho^2)^2\end{aligned}$$

Thus, $\sqrt{n}(r-\rho) = \sqrt{n-1}(f(\frac{1}{n-1}\sum_{i=2}^n V_i^2, \frac{1}{n-1}\sum_{i=2}^n W_i^2, \frac{1}{n-1}\sum_{i=2}^n V_iW_i) - f(1,1,\rho))$ has a limiting distribution which is $Normal(0,(1-\rho^2)^2)$. ■

31 A variance-stabilizing transformation (31)

Recall that $\sqrt{n}(r-\rho)$ has a limiting $Normal(0,(1-\rho^2)^2)$ distribution.

Consider $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$. Then, $\tanh^{-1}(x) = \frac{1}{2}\ln(\frac{1+x}{1-x}) = \frac{1}{2}(\ln(1+x) - \ln(1-x))$. Set $f(x) = \tanh^{-1}(x)$. Then, $f'(x) = \frac{1}{2}(\frac{1}{1+x} + \frac{1}{1-x}) = \frac{1}{1-x^2}$. Applying the Propogation of Errors Theorem, we find that $\sqrt{n}(f(r) - f(\rho))$ has a limiting normal distribution with mean $\vec{0}$ and variance $(1-\rho^2)^2(\frac{1}{1-\rho^2})^2 = 1$. Thus, $\sqrt{n}(\tanh^{-1}(r) - \tanh^{-1}(\rho))$ has a limiting standard normal distribution.

(This allows us to construct confidence intervals for $\tanh^{-1}(\rho)$ and then take the hyperbolic tangent of the endpoint to find confidence intervals for ρ.)