Mathematical Statistics II

STA2212H S LEC9101

Week 2

January 14 2025


The New York Times

Surgeon General Calls for Cancer Warnings on Alcohol

Dr. Vivek Murthy's report cites studies linking alcoholic beverages to at least seven malignancies, including breast cancer. But to add warning labels, Congress would have to act.

Listen to this article · 9:24 min Learn more

Share full article

2.1K

The advisory called for updating labels on all alcoholic beverages with a warning that

Today

1. Upcoming seminars of interest
2. Recap Jan 7 + KL-divergence + delta method
3. Likelihood ratio tests and profile likelihood SM 4.5, MS 7.4
4. computing MLEs, EM algorithm, nonparametric MLE, misspecified models MS Ch. 5.5,6,7; 3.5
5. Bayesian inference and estimation MS Ch.5.8
6. HW1, Statistics in the News

Upcoming seminars

- Department Seminar Thursday January 16 11.00 – 12.00
  Hydro Building, Room 9014
  Deanna Needell, UCLA "Fairness and Foundations in Machine Learning "
- CANSSI Ontario online
  Genevieve Gauthier, HEC "Enhancing deep hedging of options"

Recap

- data $\boldsymbol{x}_n = (x_1, \dots, x_n)$ independent observations; model $f(\boldsymbol{x}_n; \theta) = \prod f(x_i; \theta), \quad \theta \in \mathbb{R}$
- limit theorem $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, I^{-1}(\theta))$ $\quad I_n(\theta) = nI(\theta) = \text{var}_\theta\{\ell'(\theta; \boldsymbol{X}_n)\}$
- approximation $\hat{\theta} \dot{\sim} N\{\theta, I_n^{-1}(\hat{\theta})\}$, or $\hat{\theta} \dot{\sim} N\{\theta, j^{-1}(\hat{\theta})\}$ $\quad j(\theta) = -\ell''(\theta; \boldsymbol{x}_n)$

Recap

- data $\boldsymbol{x}_n = (x_1, \ldots, x_n)$ independent observations; model $f(\boldsymbol{x}_n; \theta) = \prod f(x_i; \theta)$, $\quad \theta \in \mathbb{R}$
- limit theorem $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, I^{-1}(\theta))$ $\quad I_n(\theta) = nI(\theta) = \text{var}_\theta\{\ell'(\theta; \boldsymbol{X}_n)\}$
- approximation $\hat{\theta} \,\dot{\sim}\, N\{\theta, I_n^{-1}(\hat{\theta})\}$, or $\hat{\theta} \,\dot{\sim}\, N\{\theta, j^{-1}(\hat{\theta})\}$ $\quad j(\theta) = -\ell''(\theta; \boldsymbol{x}_n)$

- data $\boldsymbol{x}_n = (x_1, \ldots, x_n)$ independent observations; model $f(\boldsymbol{x}; \boldsymbol{\theta}) = \prod f(x_i; \boldsymbol{\theta})$, $\quad \boldsymbol{\theta} \in \mathbb{R}^p$
- limit theorem $\sqrt{n}\{I(\boldsymbol{\theta})\}^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} N_p(\boldsymbol{0}, \mathcal{I}_p)$
- approximation $\hat{\boldsymbol{\theta}} \,\dot{\sim}\, N_p\{\boldsymbol{\theta}, I_n^{-1}(\hat{\boldsymbol{\theta}})\}$, or $\hat{\boldsymbol{\theta}} \,\dot{\sim}\, N_p\{\boldsymbol{\theta}, j^{-1}(\hat{\boldsymbol{\theta}})\}$ $\quad$ Check Cheatsheet

Recap

- data $\boldsymbol{x}_n = (x_1, \ldots, x_n)$ independent observations; model $f(\boldsymbol{x}_n;\theta) = \prod f(x_i;\theta)$, $\quad \theta \in \mathbb{R}$
- limit theorem $\sqrt{n}(\hat\theta - \theta) \xrightarrow{d} N(0, I^{-1}(\theta))$ $\quad I_n(\theta) = nI(\theta) = \mathrm{var}_\theta\{\ell'(\theta;\boldsymbol{X}_n)\}$
- approximation $\hat\theta \dot\sim N\{\theta, I_n^{-1}(\hat\theta)\}$, or $\hat\theta \dot\sim N\{\theta, j^{-1}(\hat\theta)\}$ $\quad j(\theta) = -\ell''(\theta;\boldsymbol{x}_n)$

- data $\boldsymbol{x}_n = (x_1, \ldots, x_n)$ independent observations; model $f(\boldsymbol{x};\boldsymbol{\theta}) = \prod f(x_i;\boldsymbol{\theta})$, $\quad \boldsymbol{\theta} \in \mathbb{R}^p$
- limit theorem $\sqrt{n}\{I(\boldsymbol{\theta})\}^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} N_p(\mathbf{0}, \mathcal{I}_p)$
- approximation $\hat{\boldsymbol{\theta}} \dot\sim N_p\{\boldsymbol{\theta}, I_n^{-1}(\hat{\boldsymbol{\theta}})\}$, or $\hat{\boldsymbol{\theta}} \dot\sim N_p\{\boldsymbol{\theta}, j^{-1}(\hat{\boldsymbol{\theta}})\}$ $\quad$ Check Cheatsheet

- Theorem 5.4
- data $\boldsymbol{x}_n = (x_1, \ldots, x_n)$ independent observations $\quad J(\theta) = \mathrm{E}_\theta\{-\ell''(\theta;\boldsymbol{X})\}/n$
- limit theorem $\sqrt{n}(\hat{\boldsymbol{\theta}} - \theta) \xrightarrow{d} N\{\mathbf{0}, J^{-1}(\theta)I(\theta)J^{-1}(\theta)\}$ $\quad$ slightly more general
- In MS Examples 5.14 and 5.15, $I(\theta) = J(\theta)$

... Recap

- proof requires many smoothness conditions on underlying model
- proof requires $\hat{\theta} \xrightarrow{p} \theta$ MS p.249; Thm 5.1,2

- i.i.d. can often be weakened to independent (not i.d.) observations, or even dependent need WLLN and CLT

- MS Theorem 5.3, p.253 has a careful proof for $\theta \in \mathbb{R}$

see also likelihood cheatsheet long version

- key step is

$$\sqrt{n}(\hat{\theta} - \theta) \simeq \frac{-n^{-1/2}\Sigma_{i=1}^{n}\ell'(X_i;\theta)}{n^{-1}\Sigma_{i=1}^{n}\ell''(X_i;\theta)} \quad \text{and}$$

... Recap

- proof requires many smoothness conditions on underlying model
- proof requires $\hat\theta \xrightarrow{p} \theta$ MS p.249; Thm 5.1,2

- i.i.d. can often be weakened to independent (not i.d.) observations, or even dependent need WLLN and CLT

- MS Theorem 5.3, p.253 has a careful proof for $\theta \in \mathbb{R}$

see also likelihood cheatsheet long version

- key step is

$$\sqrt{n}(\hat\theta - \theta) \simeq \frac{-n^{-1/2}\Sigma_{i=1}^n \ell'(X_i;\theta)}{n^{-1}\Sigma_{i=1}^n \ell''(X_i;\theta)} \quad \text{and}$$

- vector version is

$$\sqrt{n}\Sigma_{k=1}^p(\hat\theta_k - \theta_k)\{n^{-1}\ell''_{jk}(\hat{\boldsymbol\theta})\} \simeq -n^{-1/2}\ell'_j(\boldsymbol\theta),$$

$j = 1, \ldots, p$

- maximum likelihood estimators minimize the KL-divergence to the data
- KL divergence from f_0 true to f_θ model :

$$KL(f_\theta; f_0) \equiv \mathrm{E}_{f_0} \log\left\{\frac{f_0(X)}{f_\theta(X)}\right\} = -\mathrm{E}_{f_0} \log\{f(X;\theta)\} + \mathrm{E}_{f_0} \log f_0(X)$$

- estimate of $\mathrm{E}_{f_0} \log\{f(X;\theta)\}$?

- minimize $KL(f_\theta; f_0)$ same as maximize $\ell(\theta; x_1, \ldots, x_n)$

Your friend the delta-method

Suppose $\theta \in \mathbb{R}^p$, $\boldsymbol{X_n} = (X_{1n}, \ldots, X_{pn}) \in \mathbb{R}^p$

$$a_n(\boldsymbol{X_n} - \theta) \xrightarrow{d} \boldsymbol{Z},$$

and $g(\boldsymbol{x})$ is continuously differentiable at θ, then $\{g_1(\boldsymbol{x}), \ldots g_k(\boldsymbol{x})\} \in \mathbb{R}^k$

$$a_n\{g(\boldsymbol{X_n}) - g(\theta)\} \xrightarrow{d} D(\theta)\boldsymbol{Z}$$

where $D(\theta) =$

... Your friend the delta-method

MS Th.3.4 and p.148

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{d} N_p\{0, I^{-1}(\theta)\}$$

$$\sqrt{n}\{g(\widehat{\theta}_n) - g(\theta)\} \xrightarrow{d} N\{0, g'(\theta)^T I^{-1}(\theta) g'(\theta)\}$$

See also AoS §9.9

Example

$$X_1, \ldots, X_n \text{ i.i.d. Gamma } (\alpha, \lambda)$$

$$f(x_i; \lambda, \alpha) = \frac{1}{\Gamma(\alpha)} \lambda^{\alpha} x_i^{\alpha - 1} \exp(-\lambda x_i)$$

… Example

find a.var($\hat{\mu}$) via mv delta method

Newton-Raphson:

$$0 = \ell'(\hat\theta) \approx \ell'(\theta_0) + \ell''(\theta_0)(\hat\theta - \theta_0)$$
$$\hat\theta \approx \theta_0 - \{\ell''(\theta_0)\}^{-1}\ell'(\theta_0)$$

- suggests iteration

$$\hat\theta^{(k+1)} = \hat\theta^{(k)} + \{-\ell''(\hat\theta^{(k)})\}^{-1}\ell'(\hat\theta^{(k)}) = \quad \hat\theta^{(k)} + \frac{S(\hat\theta^{(k)})}{H(\hat\theta^{(k)})}$$

MS p.270; note change in notation

- requires reasonably good starting values for convergence
- need $-\ell''(\hat\theta^{(k)})$ to be non-negative definite
- Fisher scoring replaces $-\ell''(\cdot)$ by its expected value
- N-R and F-S are gradient methods; many improvements have been developed
- solution is a global max only if $\ell(\theta)$ is concave

E-M algorithm: procedure

- complete data $\boldsymbol{X} \sim f_{\boldsymbol{X}}(\boldsymbol{x}; \theta)$
- observed data $y = (y_1, \ldots, y_m)$, with $y_i = g_i(\boldsymbol{x})$ many-to-one
- joint density $f_Y(y; \theta) = \int_{A(y)} f_{\boldsymbol{X}}(\boldsymbol{x}; \theta) d\boldsymbol{x}$ $A(y) = \{\boldsymbol{x}; y_i = g_i(\boldsymbol{x}), i = 1, \ldots, m\}$
- algorithm:
  1. (E step) estimate the complete data log-likelihood function for θ using current guess $\hat{\theta}^{(k)}$
  2. (M step) maximize that function over θ and update to $\hat{\theta}^{(k+1)}$ usually by N-R or Fisher scoring
- likelihood function increases at each step
- can be implemented in complex models
- doesn't automatically provide an estimate of the asymptotic variance

but methods exist to obtain this as a side-product

Max vs Sup

DEFINITION. Suppose that for a sample $\boldsymbol{x} = (x_1, \cdots, x_n)$, $L(\theta)$ is maximized (over Θ) at $\theta = S(\boldsymbol{x})$:

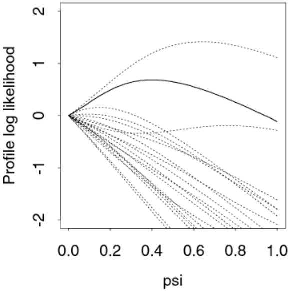
$$\sup_{\theta \in \Theta} \mathcal{L}(\theta) = \mathcal{L}(S(\boldsymbol{x}))$$

(with $S(\boldsymbol{x}) \in \Theta$). Then the statistic $\widehat{\theta} = S(\boldsymbol{X})$ is called the maximum likelihood estimator (MLE) of θ. ($S(\boldsymbol{x})$ is sometimes called the maximum likelihood estimate based on $\boldsymbol{x}$.)

… Max vs Sup

4.6 · *Non-Regular Models*

Figure 4.8 Likelihood inference for t_ν distribution. Left: profile log likelihoods for $\psi = \nu^{-1}$ for maize data (solid), and for 19 simulated normal samples (dots); $\psi = 0$ corresponds to the $N(\mu, \sigma^2)$ density. Right: χ_1^2 probability plot for the 1237 positive values of the likelihood ratio statistic $W_p(0)$ observed in 5000 simulated normal samples of size 15; the rest had $W_p(0) = 0$.



Example

MS Ex.5.25

- $f_X(x_i; \lambda, \mu, \alpha) = \alpha \frac{e^{-\lambda}\lambda^x}{x!} + (1-\alpha)\frac{e^{-\mu}\mu^x}{x!}, \quad x = 1, 2, ...; \lambda, \mu > 0, 0 < \alpha < 1$
- Observed data: $x_1, \ldots, x_n$
- Complete data: $(x_1, y_1), \ldots, (x_n, y_n)$; $y_i \sim \textit{Bernoulli}(\alpha)$
- Complete data log-likelihood function:

$$\ell_c(\alpha, \lambda, \mu; y, x) = \sum_{i=1}^{n} y_i\{\log(\alpha) + x_i \log(\lambda) - \lambda\} + \sum_{i=1}^{n}(1 - y_i)\{\log(1-\alpha) + x_i \log(\mu) - \mu\}$$

-

$$\mathrm{E}_{\hat{\theta}^{(k)}}\{\ell_c(\alpha, \lambda, \mu; y, x) \mid x\} = \sum_{i=1}^{n} \hat{y}_i\{\log(\alpha)+x_i \log(\lambda)-\lambda\}+\sum_{i=1}^{n}(1-\hat{y}_i)\{\log(1-\alpha)+x_i \log(\mu)-\mu\}$$

- $\hat{y}_i = \mathrm{E}(Y_i \mid x_i; \hat{\theta}^{(k)})$ see p.280 for exact value
- maximizing values of $\theta = (\alpha, \lambda, \mu)$ can be obtained in closed form p.281

... Example

Likelihood ratio statistic

- model $f(\boldsymbol{x};\boldsymbol{\theta}), \quad \theta \in \mathbb{R}^p$
- likelihood and log-likelihood function $L(\boldsymbol{\theta};\boldsymbol{x}), \quad \ell(\boldsymbol{\theta};\boldsymbol{x})$
- maximum likelihood estimator $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\boldsymbol{x})$

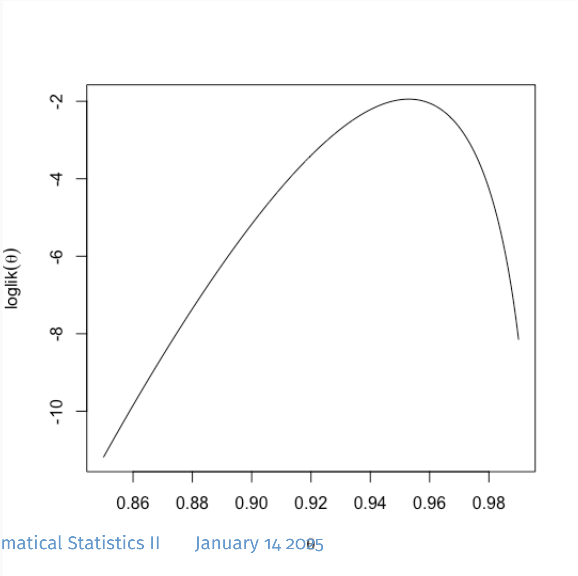
- hypothesized value $\boldsymbol{\theta}_0$ for $\boldsymbol{\theta}$
- likelihood ratio statistic $\quad w(\boldsymbol{\theta}_0) = 2\{\ell(\widehat{\boldsymbol{\theta}}) - \ell(\boldsymbol{\theta}_0)\}$

- Theorem: Under ... regularity conditions on the model ... if $\boldsymbol{\theta}_0$ is the true value

$$w(\boldsymbol{\theta}_0) \xrightarrow{d} \chi^2_p, \quad n \to \infty,$$

- Approximation: $\{\boldsymbol{\theta} : w(\theta) \geq \chi^2_p(\alpha)/2\}$ is a $1-\alpha$ confidence set for $\boldsymbol{\theta}$

$\mathrm{pr}\{\chi^2_p \geq \chi^2_p(\alpha)\} = \alpha$



loglik(θ)
-2
-4
-6
-8
-10
0.86
0.88
0.90
0.92
0.94
0.96
0.98
θ

... Likelihood ratio statistic

likelihood ratio statistic $w(\theta_0) = 2\{\ell(\widehat{\theta}) - \ell(\theta_0)\} \xrightarrow{d} \chi^2_p$

... Likelihood ratio statistic

likelihood ratio statistic $\quad w(\theta_0) = 2\{\ell(\widehat{\theta}) - \ell(\theta_0)\} \xrightarrow{d} \chi^2_p$

Aside: profile version

Nonparametric MLE

- sample $x_1, \ldots, x_n$ independent, identically distributed, with cdf F

no parametric model assumed

- likelihood function $L(F) = \prod f(x_i)$

- assume solution puts mass only at $x_1, \ldots, x_n$
- log-likelihood function $\ell(p) = \sum_{i=1}^{n} \log(p_i)$

Nonparametric MLE

- sample $x_1, \dots, x_n$ independent, identically distributed, with cdf F

  no parametric model assumed
- likelihood function $L(F) = \prod f(x_i)$
- assume solution puts mass only at $x_1, \dots, x_n$
- log-likelihood function $\ell(p) = \sum_{i=1}^n \log(p_i)$
- maximized at $p_i = 1/n, i = 1, \dots, n$

  Lagrange
- gives empirical cdf

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n 1(X_i \le x)$$

- plug-in principle: if $\theta = T(F), \hat{\theta} = T(\hat{F}_n)$

  $T(F) = \int h(x)dF(x)$, e.g.

Misspecified models

- model assumption $X_1, \dots, X_n$ i.i.d. $f(x;\theta), \theta \in \Theta$
- true distribution $X_1, \dots, X_n$ i.i.d. $F(x)$ notation
- maximum likelihood estimator based on model:

$$\sum_{i=1}^{n} \ell'(\hat{\theta}_n; X_i) = 0$$

- what is $\hat{\theta}_n$ estimating ?

Misspecified models

- model assumption $X_1, \ldots, X_n$ i.i.d. $f(x;\theta), \theta \in \Theta$
- true distribution $X_1, \ldots, X_n$ i.i.d. $F(x)$ — notation
- maximum likelihood estimator based on model:

$$\sum_{i=1}^{n} \ell'(\hat{\theta}_n; X_i) = 0$$

- what is $\hat{\theta}_n$ estimating ?
- define the parameter $\theta(F)$ by

$$\int_{-\infty}^{\infty} \ell'\{x;\theta(F)\} dF(x) = 0$$

- $$\sqrt{n}\{\hat{\theta}_n - \theta(F)\} \xrightarrow{d} N(0, \sigma^2)$$

- $$\sigma^2 = \frac{\int [\ell'\{x;\theta(F)\}]^2 dF(x)}{(\int [\ell''\{x;\theta(F)\}]^2 dF(x))^2}$$

- $$\sqrt{n}\{\hat{\theta}_n - \theta(F)\} \xrightarrow{d} N(0, \sigma^2)$$
- $$\sigma^2 = \frac{\int [\ell'\{x; \theta(F)\}]^2 dF(x)}{(\int [\ell''\{x; \theta(F)\}]^2 dF(x))^2}$$
- more generally, for $\theta \in \mathbb{R}^p$,
$$\sqrt{n}\{\hat{\theta}_n - \theta(F)\} \xrightarrow{d} N_p\{0, G^{-1}(F)\}$$
- $$G(F) = J(F) I^{-1}(F) J(F),$$
- $$J(F) = \int -\ell''\{\theta(F); x_i\} dF(x_i), \quad I(F) = \int \{\ell'(\theta(F); x_i)\}\{\ell'(\theta(F); x_i)\}^T dF(x_i)$$

Godambe information

sandwich variance

Misspecified models

MS 5.5

Bayesian estimation

model

prior

posterior

sample

Frequentist and Bayesian contrast

Frequentist:

- There is a fixed parameter (unknown) we are trying to learn
- Our methods are evaluated using probabilities based on $f(x;\theta)$

Bayesian:

- The parameter can be treated as a random variable
- We model its distribution $\pi(\theta)$
- Combine this with a model $f(x \mid \theta)$
- Update prior belief on the basis of the data

Example: Binomial

$X_1, \ldots, X_n$ i.i.d. Bernoulli (θ)

$$\pi(\theta; \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}, 0 < \theta < 1$$

posterior mean, mode

Statistics in the News

The New York Times

Surgeon General Calls for Cancer Warnings on Alcohol

Dr. Vivek Murthy's report cites studies linking alcoholic beverages to at least seven malignancies, including breast cancer. But to add warning labels, Congress would have to act.

Listen to this article · 9:24 min Learn more | Share full article | 2.1K

The advisory called for updating labels on all alcoholic beverages with a warning that drinking heightens the risk for at least seven cancers, including common ones like breast and colon cancers. Ruth Fremson/The New York Times

- "For decades, moderate drinking was said to help prevent heart attacks and strokes."
- "But growing research has linked drinking, sometimes even within the recommended limits, to various types of cancer"
- "But alcohol directly contributes to 100,000 cancer cases and 20,000 related deaths each year, the surgeon general, Dr. Vivek Murthy, said.
- He called for updating the labels to include a heightened risk of breast cancer, colon cancer and at least five other malignancies now linked by scientific studies to alcohol consumption."
- "The current warning label has not been changed since it was adopted in 1988, even though the link between alcohol and breast cancer has been known for decades."

NY Times

NATIONAL ACADEMIES *Sciences Engineering Medicine*

NATIONAL ACADEMIES PRESS
Washington, DC

This PDF is available at http://nap.nationalacademies.org/28582

Review of Evidence on Alcohol and Health (2025)

DETAILS

254 pages | 6 x 9 | PAPERBACK
ISBN 978-0-309-73115-7 | DOI 10.17226/28582

CONTRIBUTORS

Bruce N. Calonge and Katrina Baum Stone, Editors; Committee on Review of Evidence on Alcohol and Health; Food and Nutrition Board; Health and Medicine Division; National Academies of Sciences, Engineering, and Medicine

BUY THIS BOOK

Drinking less is better

We now know that even a small amount of alcohol can be damaging to health.

Science is evolving, and the recommendations about alcohol use need to change.

Research shows that no amount or kind of alcohol is good for your health.
It doesn't matter what kind of alcohol it is—wine, beer, cider or spirits.

Drinking alcohol, even a small amount, is damaging to everyone, regardless of age, sex, gender, ethnicity, tolerance for alcohol or lifestyle.

That's why if you drink, it's better to drink less.

Alcohol consumption per week

Drinking alcohol has negative consequences. The more alcohol you drink per week, the more the consequences add up.

0 drinks per week Not drinking has benefits, such as better health, and better sleep.	No risk	0
1 to 2 standard drinks per week You will likely avoid alcohol-related consequences for yourself and others.	Low risk	1, 2
3 to 6 standard drinks per week Your risk of developing several different types of cancer, including breast and colon cancer, increases.	Moderate risk	3, 4, 5, 6
7 or more standard drinks per week Your risk of heart disease or stroke increases. **Each additional standard drink** Radically increases the risk of these alcohol-related consequences.	Increasingly high risk	7, 8, + ++

During pregnancy, none is the only safe option.

A standard drink means:

Beer
341 ml (12 oz) of beer
5% alcohol

or

Cooler, cider, ready-to-drink
341 ml (12 oz) of drinks
5% alcohol

or

Wine
142 ml (5 oz) of wine
12% alcohol

or

Spirits
(whisky, vodka, gin, etc.)
43 ml (1.5 oz) of spirits
40% alcohol