STAT 111 Midterm Cheatsheet

Compiled by Zad Chin. Material based on Professor Joe Blitzstein's lectures and Joseph K. Blitzstein and Neil Shephard Lecture Notes for STAT 111.

Key STAT 111 terms

In the classic inference problem, we start by considering the i.i.d. observations

$$Y = (Y_1, \ldots, Y_n),$$

which are random variables representing the data, which then crystallize to

$$y = (y_1, \ldots, y_n).$$

A *statistic* is a function T of the random vector Y, and common statistics include the sample mean, sample median, sample mode, sample variance, and various quantiles of the data. We assume that the data we collect behave according to a *model*. This model is *parametric* if θ is finite-dimensional and *nonparametric* if θ is infinite-dimensional. Then,

- An *estimand* is a quantity of interest. Example: θ.
- An *estimator* is a random variable that encapsulates the method we use to estimate the estimand. Example: $\bar{Y}$.
- An *estimate* is a number that represents the crystallized version of some constructed estimator. Example: $\bar{y}$.

Likelihoods and MLE

Likelihoods

The *likelihood* function describes the probability of observing the data. In other words, it is a function L of the estimand θ given fixed data y:

$$L(\theta) = p(y \mid \theta).$$

In the special case where $y = (y_1, \ldots, y_n)$, with the y_j's coming from i.i.d. random variables, we can factor the joint density $p(y \mid \theta) = p(y_1, \ldots, y_n \mid \theta)$ and get

$$L(\theta) = \prod_{j=1}^{n} p(y_j \mid \theta).$$

Log Likelihoods

Loglikelihood ℓ is the logarithm of the likelihood:

$$\ell(\theta) = \log L(\theta).$$

In the usual case of $y_1, \ldots, y_n$ coming from i.i.d. random variables, we find that the loglikelihood is a sum of n terms, and taking derivatives is now easy:

$$\ell(\theta) = \log \prod_{j=1}^{n} p(y_j \mid \theta) = \sum_{j=1}^{n} \log p(y_j \mid \theta).$$

Finding the MLE

The steps you can carry out to find the MLE of θ given the data y:

1. Find $L(\theta)$ and take logs to find $\ell(\theta)$.
2. Find $\ell'(\theta)$, set it to zero, and solve for θ (call this $\hat{\theta}$).
3. Find $\ell''(\theta)$ and check that $\ell''(\hat{\theta}) < 0$.
4. With this, $\hat{\theta}$ is your maximum likelihood estimate!
5. To find the maximum likelihood estimator, convert the y_j's into Y_j's.

Reparameterization

Likelihood Invariance Allow θ to be an estimand of interest, and let $\psi = g(\theta)$, where g is injective. Then, $L(\psi) = L(\theta)$.
MLE Invariance If g is injective and $\psi = g(\theta)$, the MLE of ψ is equal to the MLE of θ evaluated at g. After all, maximizing $L(\psi)$ is equivalent to maximizing $L(\theta)$ because applying g is an inequality preserving operation.

Score Function

The *score* function is defined to be the first derivative of the loglikelihood:

$$s(\theta) = \ell'(\theta).$$

To find the MLE, we set $s(\theta) = 0$ and solve for θ, which we call $\hat{\theta}$.

Information Inequality

Let $Y = (Y_1, \ldots, Y_n)$ be a random vector of i.i.d. random variables, and suppose that the following regularity conditions hold:

- The support of Y does not depend on θ.
- All expectations and derivatives exist.

Then, both equalities below hold; the latter is known as the information equality.

$$\mathbb{E}s(\theta) = 0, \quad \mathbb{V}s(\theta) = -\mathbb{E}s'(\theta).$$

Fisher Information

The *Fisher information* for a parameter θ is defined as $\mathcal{I}_Y(\theta) = \mathbb{V}s(\theta)$
Remarks:

- Letting $\mathcal{J}_Y(\theta)$ denote $-\mathbb{E}s'(\theta)$, we have $\mathcal{I}_Y(\theta) = \mathcal{J}_Y(\theta)$ (the information equality!) if the regularity conditions mentioned earlier hold.
- You might sometimes see θ^* used instead of θ to really emphasize the fact that the entry to $\mathcal{I}_Y$ is the "true value" of θ.
- Be wary not to confuse $\mathcal{I}_Y(\theta)$ and $\mathcal{I}_{Y_1}(\theta)$. The former is the Fisher information with respect to the entire data vector $Y = (Y_1, \ldots, Y_n)$, while the latter is the Fisher information with respect to the single observation Y_1.
- In fact, if our random variables $Y_1, \ldots, Y_n$ are i.i.d., we have $\mathcal{I}_Y(\theta) = n\mathcal{I}_{Y_1}(\theta)$.

Fisher under Reparameterization: Suppose that $\tau = g(\theta)$, where g is injective and differentiable. Then, we have the relation $\mathcal{I}_Y(\tau) = \mathcal{I}_Y(\theta)/g'(\theta)^2$.

Methods of Moments

Finding MoM

Let $Y_1, \ldots, Y_n$ be i.i.d. random variables. Recall that we can freely use the notation

$$\bar{Y} = \frac{1}{n}\sum_{j=1}^{n} Y_j.$$

In a similar manner, for any kth moment, we can, without rederivation, notate

$$\bar{Y}^k = \frac{1}{n}\sum_{j=1}^{n} Y_j^k.$$

Then, the *method of moments* (MoM) estimator for some parameter θ is found by:

1. Replacing each component of θ with a corresponding component of $\hat{\theta}$.
2. Replacing the first $\dim\theta$ moments $\mathbb{E}Y_1^k$ from the model with $\bar{Y}^k$.

Finally, one can solve for each component of $\hat{\theta}$ in terms of sample moments.

Properties of MoM

If $\mathbb{V}Y_1^k < \infty$, $\mathbb{E}\bar{Y}^k = \mathbb{E}Y_1^k$ and $\mathbb{V}\bar{Y}^k = \mathbb{V}(Y_1^k)/n$. Moreover, by the CLT, we obtain

$$\sqrt{n}(\bar{Y}^k - \mathbb{E}Y_1^k) \to_{\mathcal{D}} \mathcal{N}(0, \mathbb{V}Y_1^k).$$

Note that, in general, $\mathbb{E}\hat{\theta} \neq \theta$ if $\hat{\theta}$ is an MoM estimator, even though $\mathbb{E}\bar{Y}^k = \mathbb{E}Y_1^k$

Bias, Standard Error and Loss

Bias

The *bias* of an estimator $\hat{\theta}$ for an estimand θ is $\mathrm{Bias}(\hat{\theta}) = \mathbb{E}\hat{\theta} - \theta$. Bias describes the average difference of an estimator from the estimand, and *not* the error of any particular estimate.

Standard Error

The *standard error* of an estimator $\hat{\theta}$ is $\mathrm{SE}(\hat{\theta}) = (\mathbb{V}\hat{\theta})^{1/2}$. You might notice that this is the same as the standard deviation of $\hat{\theta}$.

Loss Function

A *loss function* is a function $\mathrm{Loss}(\theta, x) \geq 0$ that is assumed to be convex in x with the property that $\mathrm{Loss}(x, x) = 0$ for all x. Examples of loss functions include:

- 0-1 loss: $\mathrm{Loss}(\theta, x) = \mathbb{I}(\theta \neq x)$.
- Absolute error loss: $\mathrm{Loss}(\theta, x) = |\theta - x|$.
- Squared error loss: $\mathrm{Loss}(\theta, x) = (\theta - x)^2$.

Mean Squared Error

The expectation of the squared error loss is called the *mean squared error* (MSE):

$$\mathrm{MSE}(\theta, \hat{\theta}) = \mathbb{E}(\theta - \hat{\theta})^2.$$

Bias-Variance Decomposition

The MSE can be decomposed as

$$\mathrm{MSE}(\theta, \hat{\theta}) = \mathrm{Bias}(\hat{\theta})^2 + \mathbb{V}\hat{\theta}.$$

Nonparametric Estimators

- *Parametric estimators* like the MLE and the MoM, which only work when the underlying statistical model is parameterized by a finite number of parameters.
- *Nonparametric estimators* allow for much less structured models, and indeed, we can estimate CDFs, quantiles, and densities without appealing to the use of any parameters at all

Empirical CDF

The *empirical CDF* (ECDF) is the function $\hat{F}$ defining a random variable given by

$$\hat{F}(y) = \frac{1}{n}\sum_{j=1}^{n} 1(Y_j \leq y),$$

where we assume that i.i.d. $Y_1, \ldots, Y_n \sim F$ have been sampled. We observe that:

- $\mathbb{E}\hat{F}(y) = \mathbb{E}1(Y_1 \leq y) = F(y)$, so the ECDF is unbiased for the CDF.
- $\mathbb{V}\hat{F}(y) = F(y)(1 - F(y))/n$.
- $\hat{F}(y) \to_{\mathcal{A}} F(y)$ by the SLLN.
- $\sqrt{n}(\hat{F}(y) - F(y)) \to_{\mathcal{D}} \mathcal{N}(0, F(y)(1 - F(y)))$ by the CLT.

Sample Quantile

If $Y \sim F$, the rth *quantile* is $Q(r) = \min\{x \mid F(x) \geq r\}$. Then, for i.i.d. random variables $Y_1, \dots, Y_n$, one can consider the order statistics $Y_{(1)}, \dots, Y_{(n)}$, when the rth *sample quantile* is $Y_{\lceil nr \rceil}$. This is a nonparametric estimator for $Q(r)$.

Kernel Density Estimator

Fix some bandwidth h, and suppose that $Y_1, \dots, Y_n$ are i.i.d. with ECDF $\hat{F}$. Then,

$$\hat{p}(y) = \frac{\mathbb{I}(Y_1 \in [y - h/2, y + h/2])}{h} = \frac{\hat{F}(y + h/2) - \hat{F}(y - h/2)}{n}$$

is called the *kernel density estimator* (KDE) with respect to the r.v.s $Y_1, \dots, Y_n$.

Asymptotics

Convergence Equivalence

Fix some constant c, and allow $Y_1, \dots, Y_n$ to be random variables. Then, convergence in distribution implies convergence in probability.

Central Limit Theorem

Let $Y_1, \dots, Y_n$ be i.i.d random variables with mean $\mathbb{E}Y_1 = \mu$ and finite variance $\mathbb{V}Y_1 = \sigma^2 < \infty$. Then, the convergence

$$\frac{\sqrt{n}(\bar{Y} - \mu)}{\sigma} \to_{\mathcal{D}} \mathcal{N}(0, 1)$$

holds. Equivalently, we have $\sqrt{n}(\bar{Y} - \mu) \to_{\mathcal{D}} \mathcal{N}(0, \sigma^2)$ or $\bar{Y} \mathrel{\dot\sim} \mathcal{N}(\mu, \sigma^2/n)$.

Law of Large Numbers

Let $Y_1, \dots, Y_n$ be i.i.d. random variables with finite first moments, i.e. $\mathbb{E}|Y_1| < \infty$. Then, $\bar{Y} \to_{\mathcal{A}} \mathbb{E}Y_1$ and $\bar{Y} \to_{\mathcal{P}} \mathbb{E}Y_1$.

Continuous Mapping Theorem

Let $g : \mathbb{R} \to \mathbb{R}$ be continuous on a set A, where $\mathbb{P}(Y \in A) = 1$. Then, we discover the following:

1. $Y_n \to_{\mathcal{A}} Y$ implies $g(Y_n) \to_{\mathcal{A}} g(Y)$.
2. $Y_n \to_{\mathcal{P}} Y$ implies $g(Y_n) \to_{\mathcal{P}} g(Y)$.
3. $Y_n \to_{\mathcal{D}} Y$ implies $g(Y_n) \to_{\mathcal{D}} g(Y)$.

Slutsky's Theorem

Suppose $X_n \to_{\mathcal{D}} X$ and $Y_n \to_{\mathcal{D}} c$, where c is a constant (recall that the latter condition is equivalent to $Y_n \to_{\mathcal{P}} c$). Then,

1. $X_n + Y_n \to_{\mathcal{D}} X + c$.
2. $X_n Y_n \to_{\mathcal{D}} cX$.
3. For $c \neq 0$, $X_n / Y_n \to_{\mathcal{D}} X/c$.

Delta Methods

Suppose that $\sqrt{n}(\hat{\theta} - \theta) \to_{\mathcal{D}} \mathcal{N}(0, \omega^2)$, and let g be a function that is continuously differentiable in a neighborhood of θ. Then,

$$\sqrt{n}(g(\hat{\theta}) - g(\theta)) \to_{\mathcal{D}} \mathcal{N}(0, g'(\theta)^2 \omega^2).$$

Kullback-Leibler Divergence

The *Kullback-Leibler divergence* (also called the KL divergence or relative entropy) from a distribution defined by p to a distribution defined by q (densities) is

$$D(p, q) = \mathbb{E} \log \frac{p(Y)}{q(Y)} = \int_{-\infty}^{\infty} p(y) \log \frac{p(y)}{q(y)}\, dy,$$

where we suppose that $Y \sim p$.

Consistency of Estimators

An estimator $\hat{\theta}$ is said to be *consistent* for the estimand θ if the convergence

$$\hat{\theta} \to_{\mathcal{P}} \theta$$

holds; that is, if $\hat{\theta}$ converges in probability to the true value of the estimand.

Proving Consistency To show that some $\hat{\theta}$ is consistent for a corresponding θ:

1. Show that $\mathrm{MSE}(\hat{\theta}, \theta) \to 0$.
2. Recognize that $\theta = \mathbb{E}Y_1$ and $\hat{\theta} = \bar{Y}$ for some i.i.d. $Y_1, \dots, Y_n$, when $\hat{\theta} \to_{\mathcal{P}} \theta$ follows immediately from the WLLN.
3. Recognize that $\theta = g(\mathbb{E}Y_1)$ and $\hat{\theta} = g(\bar{Y})$ for some continuous function g and i.i.d. $Y_1, \dots, Y_n$, when $\hat{\theta} \to_{\mathcal{P}} \theta$ follows from the WLLN and CMT.
4. Fix some $\epsilon > 0$, and show that $\mathbb{P}(|\hat{\theta} - \theta| > \epsilon) \to 0$ directly.

Cramer Rao Lowe Bound

CRLB for Unbiased Estimators Under the regularity conditions mentioned at the start, if $\hat{\theta}$ is unbiased for θ,

$$\mathbb{V}\hat{\theta} \geq \frac{1}{\mathcal{I}_Y(\theta)}.$$

CRLB in the General Case If we make no assumptions about the bias of $\hat{\theta}$ for θ, we instead obtain the bound

$$\mathbb{V}\hat{\theta} \geq \frac{g'(\theta)^2}{\mathcal{I}_Y(\theta)},$$

where $g(\theta) = \mathbb{E}\hat{\theta}$. In the zero bias case, it was the case that $g(\theta) = \theta$, so $g'(\theta) = 1$.

Rao-Blackwellization

Let T be a sufficient statistic and $\hat{\theta}$ be any estimator (in both cases with respect to the estimand θ). Then, setting

$$\hat{\theta}_{RB} = \mathbb{E}(\hat{\theta} \mid T),$$

it is the case that the inequality of MSEs given by $\mathrm{MSE}(\hat{\theta}_{RB}, \theta) \leq \mathrm{MSE}(\hat{\theta}, \theta)$ holds.

Remarks

- To Rao-Blackwellize an estimator, one must condition on sufficient statistics for the theorem to hold. The theorem fails for arbitrary statistics.
- A Rao-Blackwellized estimator will have the same bias but may have an improved (smaller variance). This follows from Adam's law and Eve's law.
- Rao-Blackwellization will not change an estimator if it was already a function of the sufficient statistic T in the first place. This follows directly from the "taking out what's known" property of conditional expectation.
- In particular, Rao-Blackwellization will not improve the MLE because the MLE is always a function of the sufficient statistics as we've seen above.
- To find the Rao-Blackwell estimator, you usually need to determine conditional distributions of the form $Y \mid T$. In Statistics 111, this is usually done by citing a Statistics 110 story, so do make a relevant list of those!

Interval Estimation

Confidence Interval

An interval estimator with a coverage probability at least $1 - \alpha$ for all possible values of θ is called a $100(1 - \alpha)\%$ *confidence interval* (CI). We say that $1 - \alpha$ is the *level* of our CI, and that $(U(Y) - L(Y))/2$ is the *margin of error* of our CI.

Exact CIs

A *pivot* is a random variable that is free of any parameters (e.g. $\mathcal{N}(0, 1)$, $\mathrm{Unif}(6, 9)$). Suppose we want to build a 95% CI for θ, where we observe $Y \sim \mathcal{N}(\theta, \sigma^2)$, where σ^2 is known, say equal to 4, the strategy is as follow:

1. Do some algebraic manipulation to get a pivot. In our example, we get
$$\frac{Y - \theta}{\sigma} \sim \mathcal{N}(0, 1).$$
2. Find the values of the quantile function of the pivot evaluated at 0.025 and 0.975. In our example, $Q_{\mathcal{N}(0,1)}(0.025) \approx -1.96$ and $Q_{\mathcal{N}(0,1)}(0.975) \approx 1.96$.
3. With these values, rest assured that a 95% CI will surface from the inequality you get when you set the 0.025 quantile value as the lower bound and the 0.975 quantile as the upper bound. In our example, a 95% CI starts from
$$Q_{\mathcal{N}(0,1)}(0.025) \leq \frac{y - \theta}{\sigma} \leq Q_{\mathcal{N}(0,1)}(0.975),$$
where y is the observed value (in an experiment) of the random variable Y.
4. Rearrange this inequality to get just the parameter of interest in the middle. In our example, some rearrangement yields the following inequality:
$$y - Q_{\mathcal{N}(0,1)}(0.975)\sigma \leq \theta \leq y - Q_{\mathcal{N}(0,1)}(0.025)\sigma.$$
5. Plug in all the values you know and assert that a 95% CI has been found! In our example, suppose we observe $y = 1$. Plugging in known values,
$$1 - (1.96)(2) \leq \theta \leq 1 + (1.96)(2),$$
so we can conclude that a 95% CI for θ is given by $[-2.92, 4.92]$.

Asymptotic CIs

Sometimes, finding a suitable pivot is hard. In these cases, we appeal to the CLT, which *always* allows us to find an asymptotic $\mathcal{N}(0, 1)$ pivot. When this pivot is not nice enough, we can further improve it by using the delta method, the CMT, or Slutsky's theorem. With this, we can construct a CI as we did before, with the caveat that the resulting interval is not exact for finite n.

A useful shortcut Finally, we look at the asymptotic 95% CI that you will end up using 95% of the time. For i.i.d. $Y_1, \dots, Y_n \sim [\theta, \sigma^2]$, with σ^2 known, we have

$$\sqrt{n}(\bar{Y} - \theta) \to_{\mathcal{D}} \mathcal{N}(0, \sigma^2)$$

by the CLT. Then, an asymptotic 95% CI for θ (which you can just quote) is

$$\left[\bar{y} - \frac{1.96\sigma}{\sqrt{n}}, \bar{y} + \frac{1.96\sigma}{\sqrt{n}}\right],$$

where $\bar{y}$ is just the crystallized version of the sample mean random variable $\bar{Y}$.

Sufficiency and Factorization

Sufficient Statistics

Let $Y = (Y_1, \ldots, Y_n)$ be a sample from some model. A statistic $T(Y)$ is *sufficient* for θ if the conditional distribution of $Y \mid T$ does not depend on θ.

Factorization Criterion

The statistic $T(Y)$ is sufficient if and only if we can factor the joint density of Y as $p(y \mid \theta) = g(T(y), \theta)h(y)$.

Sufficiency of Order Statistics

Let p be *any* density parameterized by some scalar θ. Then, if $Y_1, \ldots, Y_n$ are i.i.d. with density p, it is the case that $(Y_{(1)}, \ldots, Y_{(n)})$ is sufficient for θ. After all,

$$L(\theta) \propto \prod_{j=1}^{n} p(y_j) = \prod_{j=1}^{n} p(y_{(j)}).$$

Exponential Families

Natural Exponential Families

A random variable Y follows a *natural exponential family* (NEF) if one can write

$$p(y \mid \theta) = e^{\theta y - \Psi(\theta)} h(y).$$

We call θ the natural (canonical) parameter and note that $\Psi(\theta)$ is the cumulant generating function of Y (the logarithm of the MGF of Y).

Properties of NEFs

Let Y be NEF with the canonical form defined above. Then, we can deduce that

- $\mathbb{E}Y = \Psi'(\theta)$, $\mathbb{V}Y = \Psi''(\theta)$, and the MGF $M_Y(t) = e^{\Psi(\theta+t)-\Psi(\theta)}$.
- If $Y_1, \ldots, Y_n \sim Y$ are i.i.d., $\bar{Y}$ is a sufficient statistic for θ.
- The MLE of $\mu = \mathbb{E}Y$ is $\hat{\mu} = \bar{Y}$.
- The Fisher information is $\mathcal{I}_Y(\theta) = \Psi''(\theta)$.
- We call the process where we fix $h(y)$ as a baseline distribution from which we construct the entire NEF *exponential tilting*.
- Some examples of NEFs include the Normal (with σ^2 known), the Poisson, the Binomial (with n fixed), the Negative Binomial (with r fixed), and last, but not least, the Gamma (with a known).

Exponential Families

A random variable Y follows an *exponential family* (EF) if one can write

$$p(y \mid \theta) = e^{\theta T(y) - \Psi(\theta)} h(y).$$

Some examples of EFs include the Weibull, the Normal (but now with μ known and σ^2 unknown), and the Normal (with both μ and σ^2 unknown). And to prove that a distribution follows a NEF or an EF, you need to manipulate a given density and pattern match to a general functional form (as when you find sufficient statistics).

Mathematical Tools

Taylor Approximation

First order Taylor expansion gives a linear approximation of a function g near some point x_0 as

$$g(x) \approx g(x_0) + \frac{\partial g(x_0)}{\partial x}(x - x_0).$$

For a fixed x_0, the Taylor expansion is linear in x. This approximation should be reasonably accurate when x is close to x_0.

Differentiation under the internal sign

For any function f, by DuThIS, we have that

$$\frac{d}{dx}\int_a^b f(x,t)dt = \int_a^b \frac{d}{dx} f(x,t)dt$$

Sum of Squares Identity

Now, let's talk about some additional notation that might pop up here and there. Let $Y_1, \cdots, Y_n$ be random variables. The *sample mean*, $\bar{Y}$, is the random variable

$$\bar{Y} = \frac{1}{n}\sum_{j=1}^{n} Y_j.$$

On the other hand, the *sample variance*, S^2, is the random variable given by

$$S^2 = \frac{1}{n-1}\sum_{j=1}^{n}(Y_j - \bar{Y})^2.$$

When $Y_1, \ldots, Y_n$ crystallize into the numbers $y_1, \ldots, y_n$, we can analogously define

$$\bar{y} = \frac{1}{n}\sum_{j=1}^{n} y_j, \quad s^2 = \frac{1}{n-1}\sum_{j=1}^{n}(y_j - \bar{y})^2.$$

You are encouraged to use the expressions $\bar{Y}$ and S^2, along with their crystallized analogues $\bar{y}$ and s^2, freely without having to rederive them! Now, we obtain

$$\sum_{j=1}^{n}(Y_j - c)^2 = (n-1)S^2 + n(\bar{Y} - c)^2$$

for all $c \in \mathbb{R}$! This turns out to be a really important identity that appears all the time in statistics e.g. when deriving the posterior when the prior is Uniform on $(\mu, \log\sigma)$ and the data is Normal.

Important Examples

MLE and MoM for Normal Model

Normal with known variance Let $Y_1, \cdots, Y_n$ be iid $N(\mu, \sigma^2)$ with $\theta = \mu$ unknown but σ^2 is known. The likelihood function, dropping normalizing constant is

$$L(\mu; y) = \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(y_j - \mu)^2\right\}$$

and the log-likelihood is

$$\ell(\mu; \mathbf{y}) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_j - \mu)^2 = -\frac{1}{2\sigma^2}\left\{\sum_{j=1}^{n}(y_j - \bar{y})^2 + n(\bar{y} - \mu)^2\right\}$$

It is easy to maximize $\ell(\mu; \mathbf{y})$, just set $\mu = \bar{y}$, and we observe that

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

and so $\hat{\mu}$ is unbias with standard error

$$\mathrm{SE}(\hat{\mu}) = \frac{\sigma}{\sqrt{n}}$$

Nomral with both parameters unknown Let $Y_1, \cdots, Y_n$ be iid $N(\mu, \sigma^2)$ with both parameters unknown. We will parameterize the model in terms of the mean and standard deviation, $\theta = (\mu, \sigma)$ instead of (μ, σ^2). Then, we observe that

$$L(\mu, \sigma; \mathbf{y}) = \frac{1}{\sigma^n}\exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(y_j - \mu)^2\right\}$$

and that the log likeligood is

$$\ell(\mu, \sigma; \mathbf{y}) = -\frac{1}{2\sigma^2}\left\{\sum_{j=1}^{n}(y_j - \bar{y})^2 + n(\bar{y} - \mu)^2\right\} - n\log\sigma$$

By multivariate calculus derivation (which we will skip here), we have the MLE as

$$\hat{\mu} = \bar{Y}, \hat{\sigma} = \frac{1}{n}\sum_{j=1}^{n}(Y_j - \bar{Y})^2$$

German Tank Problem

Allies capture n tanks from the Germans, estimate the number of tanks. Observe serial #'s $y_1, \cdots, y_n$, n tanks caputred all equally likely out of t total tanks.
Let's assume simple random sampling, then we have that

$$L(t) = \begin{cases} \frac{1}{\binom{t}{n}}, \text{for } y_1, \cdots, y_m \in \{1, \cdots, t\} \\ 0, \text{otherwise} \end{cases}$$

In fact, we observe that

$$L(t) = \frac{1}{\binom{t}{n}} I(t \geq \max(y_1, \cdots, y_n))$$

Hence our likelihood is monotone decreasing, and MLE, $\hat{t} = y_{(n)}$. Observe that the support depends on t(our parameters), which violates the regularity conditions, which means we cannot use MLE properties. Thus, we will consider the PMF of $\hat{t} = Y_{(n)}$, which is

$$P(Y_{(n)} = m) = \frac{\binom{m-1}{n-1}}{\binom{t}{n}} \text{for} m = n, n+1, \cdots, t$$

Then,

$$E[Y_{(n)}] = \frac{1}{\binom{t}{n}}\sum_{m=n}^{t} m\binom{m-1}{n-1} = \frac{n}{n+1}(t+1)$$

by Feynman Restaurant problem in STAT 110 4.92 (pg 210) where we observe

$$E\left(\frac{n+1}{n}Y_{(n)}\right) = t+1 \implies E\left(\frac{n+1}{n}Y_{(n)} - 1\right) = t$$

Pivot based on Student-t distribution

Let the data be i.i.d $Y_1, \cdots, Y_n \sim N(\mu, \sigma^2)$ with both parameters unknown. Suppose that we want a $1-\alpha$ CI for μ. Since σ is unknown, we can replace σ by the standard deviation $\hat{\sigma}$, but then we can only have an approximate CI. Instead, let us constrcut a pivot, the *t-statistics*

$$T = \frac{\bar{Y} - \mu}{\hat{\sigma}/\sqrt{n}} = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \times \frac{\sigma}{\hat{\sigma}}$$

Table of Distributions

Distribution	PMF/PDF and Support	Expected Value	Variance	MGF
Bernoulli Bern(p)	$P(X=1)=p$ $P(X=0)=q=1-p$	p	pq	$q+pe^t$
Binomial Bin(n, p)	$P(X=k)=\binom{n}{k}p^k q^{n-k}$ $k \in \{0,1,2,\dots n\}$	np	npq	$(q+pe^t)^n$
Geometric Geom(p)	$P(X=k)=q^k p$ $k \in \{0,\ 1,\ 2,\ \dots\}$	q/p	q/p^2	$\frac{p}{1-qe^t},\ qe^t < 1$
Negative Binomial NBin(r, p)	$P(X=n)=\binom{r+n-1}{r-1}p^r q^n$ $n \in \{0,\ 1,\ 2,\ \dots\}$	rq/p	rq/p^2	$(\frac{p}{1-qe^t})^r,\ qe^t < 1$
Hypergeometric (w,b,n)	$P(X=k)=\binom{w}{k}\binom{b}{n-k}/\binom{w+b}{n}$ $k \in \{0,1,2,\dots,n\}$	$\mu=\frac{nw}{b+w}$	$\left(\frac{w+b-n}{w+b-1}\right)n\frac{\mu}{n}(1-\frac{\mu}{n})$	messy
Poisson Pois(λ)	$P(X=k)=\frac{e^{-\lambda}\lambda^k}{k!}$ $k \in \{0,\ 1,\ 2,\ \dots\}$	λ	λ	$e^{\lambda(e^t-1)}$
Uniform Unif(a, b)	$f(x)=\frac{1}{b-a}$ $x \in (a,b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$\frac{e^{tb}-e^{ta}}{t(b-a)}$
Normal (μ,σ^2)	$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/(2\sigma^2)}$ $x \in (-\infty,\infty)$	μ	σ^2	$e^{t\mu+\frac{\sigma^2t^2}{2}}$
Exponential Expo(λ)	$f(x)=\lambda e^{-\lambda x}$ $x \in (0,\infty)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$\frac{\lambda}{\lambda-t},\ t<\lambda$
Gamma Gamma(a,λ)	$f(x)=\frac{1}{\Gamma(a)}(\lambda x)^a e^{-\lambda x}\frac{1}{x}$ $x \in (0,\infty)$	$\frac{a}{\lambda}$	$\frac{a}{\lambda^2}$	$\left(\frac{\lambda}{\lambda-t}\right)^a,\ t<\lambda$
Beta Beta(a, b)	$f(x)=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1}$ $x \in (0,1)$	$\mu=\frac{a}{a+b}$	$\frac{\mu(1-\mu)}{(a+b+1)}$	messy
Log-Normal $\mathcal{LN}(\mu,\sigma^2)$	$\frac{1}{x\sigma\sqrt{2\pi}}e^{-(\log x-\mu)^2/(2\sigma^2)}$ $x \in (0,\infty)$	$\theta=e^{\mu+\sigma^2/2}$	$\theta^2(e^{\sigma^2}-1)$	doesn't exist
Chi-Square χ_n^2	$\frac{1}{2^{n/2}\Gamma(n/2)}x^{n/2-1}e^{-x/2}$ $x \in (0,\infty)$	n	$2n$	$(1-2t)^{-n/2},\ t<1/2$
Student-t t_n	$\frac{\Gamma((n+1)/2)}{\sqrt{n\pi}\Gamma(n/2)}(1+x^2/n)^{-(n+1)/2}$ $x \in (-\infty,\infty)$	0 if $n>1$	$\frac{n}{n-2}$ if $n>2$	doesn't exist